RiCL

Research in Corpus Linguistics

RiCL

8/2 (2020)

aelinco
Asociación Española de Lingüística de Corpus

RiCL 8/2 (2020)

Editors

Paula Rodríguez-Puente and Carlos Prado-Alonso

ISSN 2243-4712

https://ricl.aelinco.es/

RiCL Research in Corpus Linguistics

Official journal of aelinco Asociación Española de Lingüística de Corpus

RiCL Research in Corpus Linguistics

8/2 (2020)

Articles

Book Reviews

RiCL Research in Corpus Linguistics

Teaching acronyms to the military: A paper-based DDL approach

Yolanda Noguera-Díaz[a] – Pascual Pérez-Paredes[b]
Technical University of Cartagena[a] / Spain
University of Murcia[b] / Spain

Abstract – This research investigates the use of Data-driven learning (DDL) tasks in the teaching and learning of acronyms in a specialised corpus. Our target population is professional military staff (n=16). The researchers collected and analysed the *Salvage and Rescue of Submarines Corpus* (SAR) where the patterning of acronyms, neglected in English for Specific Purposes (ESP), plays a substantial role. Using a mixed-methods methodology, this research looked at the students' interaction with DDL, as well as at the subsequent interviews with the students. Deductive and inductive paper-based DDL tasks with concordance lines of acronyms were used with two groups of students of different rank. Both groups found the tasks challenging and showed mixed reactions towards concordance lines. While there has been a much-needed emphasis on tools and corpus methods training in DDL, we suggest that conversations with adult, professional students about the nature of instructed language learning and language patterning are absolutely essential to promote a more active learner role in DDL approaches.

Keywords – corpora; specialised discourse; Data-driven language learning; acronyms

1. INTRODUCTION

Linguistic analyses of English for Specific Purposes (henceforth ESP) registers have turned their interest towards professional practice by looking at both their academic and their specialised discourses (Bhatia *et al.* 2011). These findings have revitalised the interest of ESP professionals in the use of authentic language in language teaching (Gavioli and Aston 2001; Gavioli 2005; Boulton and Cobb 2017).

Corpora are useful tools for both increasing teachers' language awareness and improving lesson planning. Apart from revealing hidden patterns of use, they can also help ESP teachers capture the reality of professional discourse (Gavioli and Aston 2001: 238). In the language classroom, language learners seem to improve their linguistic competence (Boulton and Cobb 2017: 348) as they engage with corpus data via Data-driven learning (henceforth DDL) and language research tasks (Mishan 2004: 219).

Research in Corpus Linguistics 8/2: 1–27 (2020). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.08.02.01

DDL explores the application of corpus linguistics tools and techniques for pedagogical purposes in the classroom. However, DDL has been implemented in limited language education contexts, mainly in Higher Education (Boulton and Cobb 2017). In universities, Boulton and Cobb (2017: 379) report that DDL in ESP contexts yields a very high *d* effect size of 2.15 on average in pre/post-test designs, which underscores the impact of corpora on language learning. 'Cohen's d' is an effect size for the comparison between two means. It is widely used in meta-analysis (Plonsky and Oswald 2014: 878). The use of DDL in professional language-learning contexts outside Higher Education classrooms, however, remains largely underexplored. New materials and empirical studies for DDL are needed (Vyatkina 2020: 306).

Our focus is a professional community that has been particularly under-researched in the specialised literature: the military. Due to the dearth of English teaching materials for the military (Noguera-Díaz and Pérez-Paredes 2019: 118), we decided to explore the viability of corpus analysis and DDL in the context of a Navy School. In this research, we examined acronyms as used in a corpus of *Salvage and Rescue of Submarines* (SAR). This paper examines the use of a corpus-driven approach and a DDL pedagogic application in an ESP context for the first time in a Military Naval School. It focuses on the experience of the students who attend their specialisation course at the Spanish Navy Submarine Warfare School, and how DDL contributes to the learning of a selection of discourse features that are relevant to their practice. Our main research question is how professional Military understand the use of DDL in their process of language learning. This research question is theoretically framed and motivated by previous efforts to use DDL across different instructional contexts (Agee 2009). It seeks to shed further understanding of how to integrate language corpora (Boulton 2012) in specialised language instruction.

Section 2 of this paper reviews the roles of acronyms in specialised languages and, particularly, in the language used by the military. Section 3 describes a DDL approach in our specialised corpus while Section 4 describes this military context and the participants. Section 5 provides the research methodology. In Section 6, data analysis is described together with the explanation of some relevant findings. Finally, in Section 7 we discuss our results and possible future applications.

2. ACRONYMS AND THE MILITARY

Acronyms are considered as essential lexical units in science and technology. They embody the economy of language as well as being space-saving. Acronyms in biomedical and clinical documents are pervasive. A study conducted at the University of Minnesota involving clinical documents (Moon *et al.* 2013) from four hospitals used a small corpus to facilitate the extraction of acronyms and the creation of a guide for new and established practitioners. Jablonski (2005) compiled a dictionary of acronyms from medical books and periodicals from the *U.S. National Library of Medicine*.

Acronyms are "words formed from the initial letters of words that make up a name" (Quirk *et al.* 1985: 1182). New acronyms are freely produced on a daily basis, especially by scientists, journalists and administrators. Minkova (2001: 83) categorises blends as subtypes of acronyms while in Stockwell and Minkova (2009: 16) acronyms are a type of shortening. Plag (2003: 13) notes that blends are based on orthography and are called acronyms. Likewise, Stockwell and Minkova (2009: 16) distinguish between 'true acronyms' (e.g. ASCII), pronounced as any other word, and 'initialisms' (e.g. *FBI*), when the letters are pronounced individually. For the purpose of this study, we will use the cover term *acronym* to include both true acronyms and initialisms.

Despite the importance of acronyms in specialised discourse and their high frequency of occurrence in different disciplines, it is not unusual to see them neglected in ESP research. A case in point is Valipouri and Nassaji (2013), who rejected the study of acronyms in their corpus analysis of academic vocabulary in chemistry research articles, as they were not considered content words. Similarly, Konstantakis (2007) compiled the *Business Word List* —a corpus with texts from business English course books devised to train students for their university business studies— but acronyms were excluded from the analysis. Finally the *Academic Word List* (AWL) does not include acronyms either (Coxhead 2016).

The Navy and, more generally all military organisations, use acronyms for different purposes, such as organisational groups, projects and technology. For example, all organisational units within the Navy have an official acronym designation, i.e. HQ-LANDCOM, which stands for 'headquarters for allied land command' (Evered 1980: 135). In the NATO open-access documents, acronyms are frequently used in written joint operation planning by the *Allied Air Forces* (AAFCE), usually in glossaries and dictionaries "to ensure uniformity in the use of terms and definitions" (*DOD Dictionary of Military and Associated Terms* 1998: 3).

Acronyms in the English Military lexicon have received some scholarly attention. For instance, Malenica and Fabijanić (2013) studied the abbreviations from a dictionary of military terms. They did an orthographic and morphological classification of these abbreviations ranging from acronyms and blends to clippings and initialisms. In particular, they highlighted the importance of these shortened word forms in military discourse as a way to facilitate their use and favour complex communication protocols. As all branches of the Armed Forces do, the Navy also uses a specialised jargon that makes it quite unintelligible outside the discipline. In this jargon, acronyms play a substantial role. Navy acronym dictionaries come in different forms, ranging from traditional paper-based dictionaries (Cutler and Cutler 2005) to published books, and from classified publications to official reports issued as directives. The *Navy Tactical Reference Publication* (NTRP-1-02) is, for example, an unclassified Navy report, while the *DOD Dictionary of Military and Associated Terms* (1998) is an instance of a publication issued by a military section. This dictionary standardises the professional language of the U.S. Navy by defining the terminology, acronyms and abbreviations used in *Navy Warfare Library* (NWL) publications.

Using corpus analyses of a specialised military corpus, Noguera-Díaz and Peréz-Paredes (2019) have found that acronyms play a fundamental role as appositions in noun phrases. In fact, acronyms are the most common type of post-modifier in the *Cartagena Military Submarine Corpus* (CMSC) (e.g. 45-CMSC: *Test firing from a UK Royal Navy nuclear attack submarine (SSN) were in June 2005*). This corpus is made up of 822,755 words and comprises twelve years of curated texts published in a variety of professional magazines and journals. In the context of noun phrase modification, the most distinctive features of the register represented in CMSC are: 1) an above-average frequency of noun+noun modification, 2) low adjectival premodification, 3) heavy appositional postmodification and 4) low prepositional phrase modification.

In CMSC, appositive nouns occurred in 39% of the instances analysed. These finding challenges previous accounts about the spread and use of postmodifiers in other registers such as English news and academic language (Biber *et al.* 1999: 642), where appositive noun phrases (e.g. *Mr Trump, president*) account for about 15% of the postmodifiers. In the specialised corpus of *Salvage and Rescue of Submarines* (SAR), which is used in this research, acronyms play a substantial role. They represent 68% of the keywords in the corpus although they do not function mainly as appositive noun phrases. In SAR, they tend to be used as premodifiers in noun phrases (*SAR operation*) or as heads in noun phrases (*the DISSUB is assigned...*). Therefore, the importance of acronyms in SAR is also assumed essential by researchers for their teaching purposes. Table 1 shows the 10 most frequent acronyms in the corpus, their full forms as well as an example of use.

MOST FREQUENT ACRONYMS	
This procedure is applicable to any submarine SAR operation whether the DISSUB is assigned to NATO or not.	DISSUB Distressed Submarine
They have agreed to adhere to policies, procedures and minimum standards in SAR, for the needs of maritime and aviation safety.	SAR Salvage and Rescue
The primary means of securing the rescue system to the dedicated MOSHIP is by twist-lock fastenings.	MOSHIP Mother ship
The Surfacing Signal must be transmitted insufficient time to ensure its receipt by the SUBOPAUTH.	SUBOPAUTH Submarine Operating Authority
This principle should similarly apply in marine incidents where a Maritime RCC will be designated the responsible.	RCC Rescue Coordination Centre
Occasionally, the RCC requires the OSC to make various search decisions. Such as search pattern selection, track spacing, and individual search area.	OSC On-scene Commander
It should be used in conjunction with ATP- 57 which deals in more detail with the recovery of escapers and rescue of survivors.	ATP Army Techniques Publication
When the distress site and possible survivors have been located the SRV will do everything possible to facilitate the task of conducting the rescue operation.	SRV Safety Research Vehicle
The submarine should be ordered to dive for short periods and use her UWT and main sonar suite to search an area preferably away from the surface ships' search.	UWT Undersea Warfare Technology
Refer to the NATO Standardization Document Database for the complete list of existing reservations.	NATO North Atlantic Treaty Organization

Table 1: Most frequent acronyms in the SAR corpus

3. DDL AND SPECIALISED LANGUAGE

The linguistic analysis of ESP registers has attracted much scholarly attention (see, e.g., Bhatia *et al.* 2012), as new professional domains demand scrutiny and pedagogical attention. Corpus analysis techniques can be used in this context by language professionals in response to emerging needs. In the foreword to Crosthwaite and Cheung (2019: xiii), a corpus-based study of the language of dentistry and its teaching, Ken Hyland has noted that, through their interaction with corpus-based materials, learners "are required to think their way into their disciplines […], identifying the particular language features, discourse practices, and communicative skills of target groups."

Johns and Dudley-Evans (1991) proposed a DDL application of corpora in learning and teaching. DDL was conceptualised as a lexico-grammatical approach that used a concordancer to analyse certain patterns in texts, and which then would be used in the construction of teaching materials. Since then, a wealth of studies in the last decade has advocated the use of corpus linguistics in language education (Boulton and Cobb 2017; Pérez-Paredes 2019), but just some of them have combined corpus linguistics methods, DDL and ESP.

It is fifteen years now that Gavioli (2005) applied the use of hands-on DDL to teach disciplinary language and improve the language learning autonomous experiences of medical students. Research in this area, however, does not seem to have made much progress (Pérez-Paredes 2019). Most researchers seem to agree that, as pointed out by Crosthwaite and Cheung (2019: 20), the use of DDL exposes language learners to evidence about language that

> allows them to understand the characteristic language features involved in producing disciplinary genres of writing, thus enhancing their understanding of the complexities of literacy within their target disciplinary field.

However, how corpus-driven disciplinary knowledge is translated into pedagogy remains controversial (Pérez-Paredes 2019). What the evidence shows (Boulton and Cobb 2017; Pérez-Paredes 2019) is that it has been in English for Academic Purposes (EAP) where we have witnessed an increased interest in the use of DDL and specialised corpora (Yoon and Hirvela 2004; Lee and Swales 2006; Boulton and Pérez-Paredes 2014; Cotos 2014; Tono *et al.* 2014; Chen and Flowerdew 2018).

Very often, the focus of ESP research is academic language in the context of a specialised domain. Carter-Thomas and Chambers (2012) studied first-person pronouns in corpora of introductions to economics research articles, integrating printed DDL concordance lines as worksheets. Other research efforts have shown an overt, direct interest in pedagogical applications. Hafner and Candlin (2007) explored a selection of legal writing tasks from a legal corpus using an online concordancer and collocation tools. They developed an online resource called *Legal Analysis and Writing Skills* (LAWS) that included an online concordancer and a collocation tool. It was designed to familiarise students with corpus tools to improve their competence in writing for legal purposes. Several task-based exercises were created in a concordancing help section on the LAWS website. The results showed that students preferred the use of the concordancer to retrieve instances of usage for modelling-based legal articles over the completion of concordancing tasks.

Some uses of DDL in ESP, however, showed positive results. Maniez (2011) studied adjectival versus nominal modification in medical English in a corpus of texts published by the *European Medicines Agency* (EMEA). The election between a premodifying noun and an adjective is difficult for French native speakers. His corpus helped students make better-informed lexical choices. The researcher created this corpus as a guide when selecting the type of modification for non-native medicine ESP writers and specialised translators. Curado-Fuentes (2016) used DDL in ESP lessons with students of business and tourism. He found that the DDL group obtained better results than the control group that followed a traditional non-DDL methodology. The researcher chose texts related to economy and business from the *Corpus of Contemporary American English* (Davies 2008). The DDL students integrated hands-on concordancing of grammatical points (verb tenses) in their lessons, and reported a most positive feedback in terms of the usefulness of examining concordance lines.

However, the combination of corpora and DDL is not a panacea for ESP contexts (Boulton 2012: 281). According to Boulton (2012), what seems to be key is finding the balance between the appropriate corpus data and the integration into the learning environment, minimising the obstacles and highlighting the potential of DDL. As suggested by Crosthwaite and Cheung (2019: 20) corpora offer educators and learners target disciplinary language that students can use "to discover the key features of disciplinary language in use." Despite the benefits identified in the specialised literature (Boulton and Cobb 2017; Pérez-Paredes 2019), there is a dearth of emic studies that explore learners' engagement with DDL through qualitative methods and interviews. Pérez-Paredes and Sánchez-Hernández (2019) is an exception. Their interviews with university researchers two years after the corpus training sessions provide insights into the writing practices of researchers in the Spanish University context, and their reluctance to use corpora when writing. In our specific context, the use of the SAR specialised corpus for pedagogical purposes is the main target of our study.

In the following sections, we will discuss the context of this study and the methodology that was adopted to carry out our research.

4. CONTEXT AND PARTICIPANTS

The Spanish Submarine Flotilla was founded on February 17, 1915 when the Miranda Act was passed by King Alfonso XIII. The Spanish Submarine Flotilla is located in the city of Cartagena and provides specialised training on a wide range of areas through monographic courses. The Submarine School provides training to officers and ratings specialising in weapons engineering and warfare operations. The Submarine School develops and trains future Spanish submarine crews (officers, petty officers and master seamen). It has four main departments: Weapons, Tactics, Energy and Propulsion. All teachers are military staff except for the languages section in which they are civil members. The Flotilla Commander is also the Base's Chief and the Submarine School's Headmaster.

This study involves naval military personnel taking one-year specialisation course before joining the *Spanish Navy Submarine Force*. The school compulsory subjects range from acoustics, communications, torpedoes, first aid, tactics, data, equipment, services to salvage and rescue. The course runs every year from September to June. Intensive six-month theory courses are followed by three training months on board. Students are divided into three groups according to their military rank: sailors, petty officers and officers. Spanish submarines are currently part of the NATO Sea Guardian and E.U. Sophia operations.

Sailors have a certificate of Compulsory Secondary Education and have completed one year of military training in a military school. This course at the Submarine School is described as a specialisation course. Officers have a four-year degree in Naval and Military studies. Both groups took either the Preliminary English Test (PET) or the First Cambridge Test (FCE) upon their arrival at the School. A total amount of sixteen military students participated in this research: ten sailors and six officers. Once these students were debriefed, they provided consent following standard ethical guidelines for good research practice (*The British Association for Applied Linguistics Ethics* 2006; see Appendix 2). No Internet connection was available during the sessions for reasons of security.

The sailors' group consists of ten male students whose mother tongue is Spanish. 10% of the sailors have a B2 profile, another 10 % a C1 English profile and 80% an A2 (see Appendix 1A for demographic information).[1] These results can be aligned with the assessment methodology used by the Armed Forces. The language proficiency levels are measured by some level descriptors included in the *Standard NATO Agreement* 6001 (STANAG 2019). STANAG includes five levels which range from 1 (survival), 2 (functional), 3 (professional), 4 (expert) to 5 (highly articulate native). See Table 2 below for equivalence.

CEFR	**STANAG 6001**
A1	0 or 1
A2	1~1+ or 2 (mostly 1)
B1	1+ or 2 (mostly 2)
B2	2~2+ or 3 (mostly 3)
C1	2~2+ or 3 (mostly 3)
C2	3~3+ or 4

Table 2: CEFR/ STANAG 6001 equivalences

As far as the officers' group is concerned, it consists of one female student and five male students. Spanish is their mother tongue. They all have a B2 English level and have developed a basic command of Naval English (mainly military ship-related vocabulary) due to their previous military academic training (see Appendix 1B for demographic information).

5. Methodology

We adopted a mixed methods research methodology. Corpus linguistics exploration and pedagogic intervention was followed by a qualitative approach within an interpretive paradigm (Taber 2013) to explore the adoption and use of DDL in a professional military context.

This was a three-stage research project whose classroom intervention went on for a month. In the first stage, the SAR corpus was put together so as to extract and analyse

[1] These levels are those established by the Common European Framework for Reference (CEFR).

the features of the *Cartagena Military Submarine Corpus* (CMSC) following the guidelines in Noguera-Díaz and Peréz-Paredes (2019). We found that 69% of the 100 most frequent keywords in the corpus are acronyms (e.g. DISUBB, SAR, COMSUBMAR...). We examined the different grammatical relations and found a tendency for these words to function either as subjects (e.g. *The Argentinean DISSUB was found six months later*) or objects (e.g. *They have finally located the DISSUB*). This analysis gave us the understanding to move on to an informed selection of materials to be used in the language classroom based on the frequency of the acronyms, their syntactic roles at the clause and the phrase levels and their collocational profile. Analyses were carried out via *Sketch Engine* (Kilgarriff 2003). In the second stage of our research, students engaged with paper-based DDL activities. Finally, in a third stage, interviews were conducted to probe into the receptions and viability of DDL in a professional context.

5.1. Stage 1: Analysis of the specialised corpus

'Salvage and Rescue of Submarines' is a compulsory subject in the syllabus of the Spanish Navy Submarine Warfare School. While endorsed and curated by the Ministry of Defence, SAR publications are non-confidential and non-restricted, which made them the ideal target for our corpus. The SAR corpus consists of 18 non-classified NATO publications, including fifteen books and manuals and three journal articles. The corpus contains 37,615 types and 717,446 tokens. Some of the most important publications here are the so-called ATP-57(i) and (ii). These are manuals that address the techniques and procedures for salvage and rescue operations involving submarines. It is published by the STANAG, which defines processes, procedures, terms and conditions for common military technical procedures or equipment among the member countries of the alliance. Each NATO state ratifies a STANAG and implements it within their military system. The purpose of STANAG-compliant procedures is to provide common operational and administrative practices and logistics. Most of the specific bibliography was provided by the officer in charge of the International Submarine Escape and Rescue Liaison Office (ISMERLO) at the Submarine School. The ISMERLO Office is based in Northwood, United Kingdom. This site provides the worldwide submarine rescue coordination and information exchange.

5.2. Stage 2: Introductory workshop on DDL and paper based DDL activities

The second stage of our study examined the informants' first contact with the SAR corpus. Both groups of students received a 60-minute introduction to the corpus. The introduction sought to unlock the potential of corpus consultation and to unveil lexico-grammatical patterning. The introduction covered aspects such as collocations, colligation and keywords. The word *submarine* was chosen as an example and some concordances lines, which included noun phrase structure (determiners and modifiers), were displayed. During the session, students were asked to identify some patterns of use and were offered the opportunity to discuss difficulties and their first reaction. The following week, students were provided with worksheets with all the concordances of the two most frequent acronyms in the corpus: DISSUB and SAR. The students were asked to examine the lines following a similar procedure to that used in Thurston and Candlin (1998) and to note the type of words that tend to premodify and postmodify the acronyms. Once they shared their findings with the group, the instructor provided explicit explanation and solved doubts or inquiries. In the third week, the instructors used a smaller selection of concordance lines of the same acronyms (DISSUB and SAR) to showcase collocational and colligational behaviour and, thus, facilitate a closer examination of the contexts in which acronyms were used. Students were provided with a worksheet that included different exercises. In the first block, learners were offered a brief explanation on word order and the verb phrase in the English language. The follow-up activities in Tables 3 and 4 were conceptualised as deductive activities.

Activity 1: Underline the finite verb phrases after the acronym DISSUB and level them.
-It was three days after the DISSUB had been found, Marine Sound Signals (MSS) off.
-They are in a hard situation unless the DISSUB has underwater Morse or voice signal.
-Unless you are in a scarcity of power, the DISSUB will try to transmit continuously.
-There are not pills available, this DISSUB crew will concentrate on using masks.

Table 3: Activity 1 - SAR corpus and finite verbs

Activity 2: What are the word classes that appear frequently before DISSUB? Are they adjectives, determiners, nouns, etc.?
-At nautical miles during all DISSUB transfer evolutions. A 20-angled wedge is needed.
-With precise angled DISSUB mating. Using a combination of trim and draught.
- Hatches and portholes in the DISSUB when equalised with the RC. A CO2 scrubbing.
-Localising a DISSUB: Maximum angle 60 degrees to the horizontal plane at any ratio.

Table 4: Activity 2 - DISSUB premodification

In the second block, students completed activities 3 and 4 (see Tables 5 and 6) without any exposure to explicit declarative knowledge. Time was provided to facilitate discussions around the completion of the activities and the lexico-grammatical points raised. In the last week, further feedback on the previous lesson activities was given and semi-structured interviews were conducted.

Activity 3: Left context of the acronym. Choose one of these three adjectives for suitable gap-nuclear, simulated, atmospheric, diesel.
-Rescues from a .... DISSUB should be taken as radiological contaminated until proven otherwise.
-Monitoring the .... DISSUB internal data during the ventilation operation is crucial for the efficacy.
-Establish UWT communications between surface and .... DISSUB in accordance with scripts.
-Rescue crewmembers from a ....DISSUB carry out basic medical training scenarios.

Table 5: Activity 3 - DISSUB left context.

Activity 4: Right context of the acronym. Choose one of these nouns for suitable gap: crew, position, request, condition.
-The ship(s) nominated by the SSRA to carry stores and equipment which may be needed to sustain DISSUB's ....
-Marking the Submarine's position. It is important that the DISSUB.............. is not lost, particularly in a tideway.
-Every effort must be made to comply with the DISSUB…...... to obtain specialist advice on what might be required.
-The purpose of the divers is to orient and familiarise the rescue unit, inspect the DISSUB…...... and damage

Table 6: Activity 4 - DISSUB right context

The same paper-based DDL activities (Boulton 2010) were used in both groups on different dates following the same protocol. The acronyms DISSUB and SAR were chosen, as they are the most frequent keywords in our corpus. These acronyms function either as the subjects of a clause and display some of the properties of regular nouns, thereby being premodified by a variety of structures (e.g. adjective phrases, noun phrases or past participles), or functioning as premodifiers in noun phrases (e.g. *A SAR operation covers the whole process).* A careful analysis of the frequency of appearance and breadth of syntactic functions helped us to decide what concordance lines to select and the target forms to discuss during the activities, particularly during week 3. The overarching objective of these DDL activities was to make the students familiar with the patterning of these words, together with discovering of some of their most common pre-and-post modifiers in (authentic) contexts.

5.3. Stage three: Semi-structured interviews

We used semi-structured interviews (Gray 2015: 213) to tap into the learners' use of DDL. The interviews were conducted in Spanish and were recorded and transcribed for further analysis. The students were debriefed and were reminded that anonymity and their right to remove themselves from the research were warranted. The sequence of the questions went from students' general personal learning experience with English —as in *What is your goal as a language learner*? *How do you learn grammar and vocabulary*?— to more precise recall of their experience with DDL, as in *Describe your experience with the DISSUB concordance lines,* or *Can you focus your attention just on the right section of the line?* (see Appendix 3 for details). The students were all cooperative and eager to provide their answers openly. In total, 3.15 hours of recorded material were transcribed: two hours in the sailors' group and one hour and fifteen minutes in the officers' group.

6. Data analysis and findings

We analysed both the DDL activities completed by the informants and their interviews. The researchers categorised the answers to the four activities for each acronym (two inductive and two deductive) as correct or wrong in order to evaluate the students' understanding of the activities. We used 'theme analysis' (Gray 2015: 319) to examine the students' reactions to DDL in the interview data. Theme analysis is a widely used data reduction and analysis method that extracts themes and subthemes from textual data in order to understand how they are interrelated (Pérez-Paredes 2020).

6.1. DDL activities

Paper-based DDL involves the study of patterns by means of printed materials prepared by language teachers or researchers (Tribble and Jones 1997). Our students had no direct access to the corpus or concordance software during the activities. One of the main advantages of paper activities is that corpora insights can be shared with a wider audience, who cannot have access to computers or the Internet. In our case, the School is heavily protected against cyber-attacks, which makes it extremely difficult for students to use their own devices or for teachers to access a Wi-Fi or a LAN point. Another positive side is that, in classroom contexts where technology is not normalised

(Bax 2003), students may feel at ease with printed concordance data. This eliminates much of the challenges discussed in the literature concerning training to use a corpus (Boulton and Cobb 2017: 350).

The first activity, which is illustrated in Table 3 (see Section 5.2.), involved 1) looking at the concordance lines, 2) paying attention to the verbs which follow DISSUB and 3) underlining them. Then, the students examined the concordance lines and underlined the words that pre-modified or post-modified the acronym in order to classify them into a morphological category. Informants were also asked to consider the verb tenses which post-modified the acronym. These activities followed an introduction to tenses in verb phrases and noun phrase complexity, where the instructor used explicit declarative knowledge about the grammar of the English verbal and noun systems.

Activity 1 and 2 (see Tables 3 and 4 in Section 5.2) followed a deductive learning approach (Flowerdew 1996: 97) that was successful in both groups (95% of the answers were correct). In-depth observation of the left and right contexts helped students infer information about the syntactic nature of both acronyms. However, the results of the third and fourth activities (see Tables 5 and 6 in Section 5.2.) yielded low scores in both groups. These activities followed and inductive learning approach that seemed to be more cognitive demanding, as the students were asked to discover patterns and analogies that implied language noticing and the use of a wider range of vocabulary. For these activities, the instructor did not offer an explicit account of the grammar or the lexical properties of the noun phrases involved. Only 20% of the sailors' group answers were correct, while in the officers' group only 30% of the answers were not.

6.2. Semi structured interviews activities

Different themes emerged from the questions that were discussed during the interviews with the two groups of students. What follows provides a summary of both the themes and the reactions to those themes in the two groups. Transcriptions are presented *verbatim*.

6.2.1. Concordance lines

The students' perceptions in both groups reflect certain feeling of confusion over the concordance lines. Students felt that going through the lines was more exhausting than other activities they were more familiar with. In general, they seemed to prefer the teacher's explicit guidance and a more traditional method. By way of example, in the sailors' group, student Number 2 affirmed: "[…] the system requires a significant effort of concentration because after the fifth line you get dizzy. Sometimes the word does the same function in each sentence and you must pay attention and make a much greater effort than normal. It is very repetitive."

In the officers' group, student Number 3 said: "I see it very intuitive but very hard. It does not help me more than a direct translation of the word in my mobile or reading the English definition in a dictionary." In addition, student Number 5 added: "It reminds me of my best English dictionary with different entries of the same word."

6.2.2. English language methodology

Students claimed they preferred a more traditional teaching method, with less innovative techniques and more teacher guidance. Sailor Number 4 said: "This is a lot of time-consuming work. Sometimes it is boring. I prefer reading and applying grammar rules in the workbook. It is almost automatic and easier for me." However, officer Number 5 added: "I would like to know more about this method. It is so new and different […]. I was very concentrated in doing well the tasks."

6.2.3. Role of vocabulary in learning a foreign language

Both groups commented on the vital role of memorisation, translation and repetition. In the sailors' group, student Number 2 said: "To learn vocabulary you must already have some knowledge, a good base of the English language. I am overwhelmed by the lines."

Student Number 7 added: "I learn vocabulary copying paragraphs and writing words repeatedly. From 1 to 10, I would give vocabulary an importance of 9." However, officer Number 6 claimed the opposite, and said: "The lower your English level, the more grammar you must learn. On a grammar basis, you could add vocabulary easily through repetition, wordlists or reading."

6.2.4. Attitude towards concordance lines

Most students were overwhelmed by the accumulation of concordance lines and, at the same time, felt some frustration with the time needed to analyse the lines. Sailor Number 9 reported that: “There are many exercises for just a word. Too much time consuming for an acronym.” Student Number 8 suggested that: “Reading these lines properly requires a significant effort of concentration. I am not used to do that.”

As for the officers, Number 6 said: “It was a different experience, strange. It is the first time I see this type of approach. It is easier for me to look up this acronym in a monolingual dictionary with different entries.”

Our students experienced more difficulty in reading the target language acronyms than reading short paragraphs with familiar vocabulary. The two less advanced students found it hard to read the concordance output. Sailor Number 10 affirmed: “I feel overwhelmed with this method and at the same time, I get discouraged if too many items in the concordance lines were unknown.”

However, officer Number 3 said: “I would like to experience more with concordances as part of my language learning experience, but I would not like to substitute the traditional English lessons for entire lessons just with concordances.”

Students also expressed their interest in DDL. Sailor Number 10 said: “The last ten or fifteen minutes at the end of the class because if you put the lines at the beginning of the lesson and you don’t understand the vocabulary, you disconnect. It’s like talking about quantum physics to my sister who is ten years old.” Student sailor Number 9 added: “Yes, I think three or four activities of this type would be fine once you have already acquired some of this specific vocabulary. Doing activities with the concordance lines at the beginning of the class become tedious and scattered.” Similarly, officer Number 1 thought: “It is good as complementary exercises in class.”

7. Discussion

While the materials and activities addressed the domain and professional discourse training needs of our learners, both groups of students agreed that interpreting corpus data and reading concordances was quite challenging. The students’ success with deductive DDL tasks seemed to be counterbalanced by the somewhat less positive results in the inductive tasks. Irrespective of the orientation of the tasks, our informants felt overall motivated and curious about DDL, though they expressed mixed reactions.

The use of interviews in a mixed-methods design facilitates the situatedness of research data in ways that surveys cannot. Particularly, we were interested in understanding how a group of military professionals framed their ideas about language learning and about DDL, and how both are entwined with values, opinions and behaviour (Cohen *et al.* 2018: 285). Our study shows that these students' language learning ideology is dominated by the Grammar-Translation method, which emphasises the mastery of grammatical rules and vocabulary. This is reflected in the way our students have learnt English vocabulary along their academic life by memorising and copying wordlists. Their perception that learning acronyms through DDL is very time-consuming lends evidence to the fact that the type of student-centred discovery learning in DDL clashes with approaches where declarative knowledge is presented to students in ways that favour a lack of learner-centred understanding of lexico-grammatical patterning. This is perhaps a major obstacle for a DDL approach in instructed language learning contexts, where an emphasis on form is met by a lack of input in the foreign language. However, the specialised literature (Boulton and Cobb 2017) has tended to emphasise the obstacles of hands-on concordance as regards corpus consultation (see Pérez-Paredes *et al.* 2011, 2012; Boulton and Cobb 2017) and the interpretation of concordance lines (Pérez-Paredes *et al.* 2011; Pérez-Paredes 2019) ignoring learners' beliefs and their situatedness in a larger social group (Ushida 2005: 49) and their ideologies about language learning (Spolsky 2004: 80). The use of paper-based DDL removes the pressure to instruct learners on how to use concordance and, as a consequence, may enhance the engagement with the interpretation of concordance lines. This area requires further attention by researchers.

We have found evidence that identifying word patterning seems to be perceived as more demanding in inductive activities than in deductive activities, so this would seem a great point of departure to have conversations with students of specialised languages about the roles of language, language form, patterning and learning. A more explicit treatment of how learning happens in instructed contexts, in particular in adult professional contexts, seems relevant as suggested by some of the students during the interviews. The reactions to the use of authentic texts were largely positive and were in line with the findings in the literature (Boulton and Cobb 2017). The group of the officials was slightly more vocal about the importance of learning English using authentic texts. We note that some of the learners' criticism towards DDL in this research may be tentatively put down to lack of awareness about the lexico-grammatical nature of language, the role of frequency and other statistical properties of language.

It has been claimed that DDL at the tertiary level seems to be effective in contexts such as law, scientific writing or healthcare education (Crosthwaite and Cheung 2019: 27). Boulton and Cobb (2017) have established that it is predominantly Higher Education students that have been extensively examined in past DDL research, and that DDL instruction has a positive impact of language gains. We also know that paper-based DDL is effective: DDL has a mean *d* effect size of 1.06 in pre/post-test designs and 0.52 in control/experimental studies (Boulton and Cobb 2017: 377). What makes our study unique is that we have taken DDL to classrooms where DDL might rarely happen, so this is a first attempt at examining the uptake of paper-based DDL with a population of military personnel that will need to be probed in other similar contexts.

Despite the short contact time with DDL, we found some evidence that our informants noticed basic patterning around the acronyms selected. Boulton (2010: 534) has pointed out that the aim of researching the use of paper-based concordance lines is not to show that DDL is superior to other approaches, but rather present learners with complementary learning that can be useful in contexts with limited time available for training. Our approach implies not only a research-informed form of instruction about acronyms, but also increasing the students' knowledge about their professional register. New training initiatives are needed so as to examine longer exposure to DDL. The context in which we developed this research seems appropriate to use paper-based DDL as hands-on concordance is not possible. While Boulton and Cobb (2017) have suggested that DDL offers a way out of overemphasis on vocabulary lists and grammar exercises, our learners provided evidence that, in the context of a Grammar-Translation methodology, which emphasises the teaching of forms (Long 1991), DDL may face obstacles that go beyond the normalisation of Information and Communication Technologies (ICT) or corpora (Bax 2003; Pérez-Paredes 2019). The compilation of the SAR corpus as well as CSMC (Noguera-Díaz and Peréz-Paredes 2019) will hopefully create the conditions for the preparation of a syllabus that includes corpus findings and DDL as cornerstones for Navy submariners. Some of the learners evaluated DDL and learning acronyms through concordance lines as extremely useful and eye opening, but for most of them the use of concordance lines was not an efficient way to learn vocabulary. This finding echoes Pérez-Paredes and Sánchez-Hernández (2019), who showed that university researchers did not generally find corpora more useful than vocabulary lists or glossaries, when writing academic English. Pérez-Paredes and Sánchez-Hernández (2019: 60) argue that "learning and development are socially motivated and happen in culturally formed settings," which explains the divergence of results worldwide and the emic quality to most research design in DDL. Bridging the gap between the emic and the globalised urgency to learn English as the *de facto* language of many professionals worldwide is quite a challenge.

8. Conclusion

Our research presents some limitations. It belongs to a specific professional context that cannot be generalised to other learning contexts, both nationally or internationally. Although the number of informants is arguably small, it is representative of the military student enrolled in professional courses every year. We assume that the intervention period was quite short, but it was arguably a necessary step in considering the implementation of further corpus-based classroom work.

We like to think that this experience will give rise to the development of an integrated DDL syllabus, where learners and researchers can find themselves more at ease with both the DDL methodology and the sort of language-related insights that emerge from interacting with concordance lines.

DDL work requires substantial contact time, particularly in hands-on concordance contexts. Although more research is needed on the selection of concordances lines and activities, an integration of paper-based DDL into current methodological options may contribute to bringing together students' awareness of language patterning in professional contexts and approaches that favour a more active learner role.

References

Agee, Jane. 2009. Developing qualitative research questions: A reflective process. *International Journal of Qualitative Studies in Education* 22/4: 431–447.

Bax, Stephen. 2003. CALL–past, present and future. *System* 31/1: 13–28.

Bhatia, Vijay, Purificación Hernández-Sánchez and Pascual Pérez-Paredes eds. 2011. *Researching Specialized Languages.* Amsterdam: John Benjamins.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Boulton, Alex. 2010. Data-driven learning: Taking the computer out of the equation. *Language Learning* 60/3: 534–572.

Boulton, Alex. 2012. Corpus consultation for ESP: A review of empirical research. In Alex Boulton, Shirley Carter-Thomas and Elizabeth Rowley-Jolivet eds., 261–291.

Boulton, Alex, Shirley Carter-Thomas and Elizabeth Rowley-Jolivet eds. 2012. *Corpus-informed Research and Learning in ESP: Issues and Applications*. Amsterdam: John Benjamins

Boulton, Alex and Tom Cobb. 2017. Corpus use in language learning: A meta-analysis. *Language Learning* 67/2: 348–393.

Boulton, Alex and Pascual Pérez-Paredes. 2014. Researching uses of corpora for language teaching and learning. *ReCALL* 26/2: 121–127.

Carter-Thomas, Shirley and Angela Chambers. 2012. From text to corpus: A contrastive analysis of first person pronouns in economics article introductions in English and French. In Alex Boulton, Shirley Carter-Thomas and Elizabeth Rowley-Jolivet eds., 15–44.

Chen, Meilin and John Flowerdew. 2018. A critical review of research and practice in Data-driven learning (DDL) in the academic writing classroom. *International Journal of Corpus Linguistics* 23/3: 335–369.

Cohen, Louis, Lawrence Manion and Keith Morrison. 2018. *Research Methods in Education*. London: Taylor and Francis.

Cotos, Elena. 2014. Enhancing writing pedagogy with learner corpus data. *ReCALL* 26/2: 202–224.

Coxhead, Averil. 2016. Reflecting on Coxhead (2000): A new academic word list. *TESOL Quarterly* 50/1: 181–185.

Council of Europe. 2001. CEFR or Common European Framework of Reference for Languages: Learning, Teaching, Assessment, Cambridge: Cambridge University Press

Crosthwaite, Peter and Lisa Cheung. 2019. *Learning the Language of Dentistry: Disciplinary Corpora in the Teaching of English for Specific Academic Purposes*. Amsterdam: John Benjamins.

Cutler, Deborah W. and Thomas J. Cutler. 2005. *Dictionary of Naval Terms*. Maryland: Naval Institute Press.

Curado-Fuentes, Alejandro. 2016. Grammatical development via DDL at the upper-intermediate level in LSP contexts. In Mary Frances Litzler, Jesús García Laborda and Cristina Tejedor Martínez eds. *Beyond the Universe of Languages for Specific Purposes: The 21st Century Perspective*. Alcalá de Henares: Servicio de Publicaciones Universidad de Alcalá de Henares, 65–68.

Davies, Mark. 2008–. *The Corpus of Contemporary American English* (COCA): 520 million words, 1990-present. http://corpus.byu.edu/coca/.

DOD Dictionary of Military and Associated Terms. 1998. Washington DC: The Joint Staff.

Evered, Roger. 1980. *The Language of Organizations: The Case of the Navy*. California: Naval Postgraduate School Monterey.

Flowerdew, John. 1996. Concordancing in language learning. In Martha Pennington ed. *The Power of CALL*. Houston: Athelstan, 97–113.

Gavioli, Laura. 2005. *Exploring Corpora for ESP Learning*. Amsterdam: John Benjamins.

Gavioli, Laura and Guy Aston. 2001. Enriching reality: Language corpora in language pedagogy. *ELT Journal* 55/3: 238–246.

Gray, David E. 2013. *Doing Research in the Real World*. London: Sage Publications.

Hafner, Christoph A. and Christopher N. Candlin. 2007. Corpus tools as an affordance to learning in professional legal education. *Journal of English for Academic Purposes* 6/4: 303–318.

Jablonski, Stanley. 2005. *Jablonski's Dictionary of Medical Acronyms and Abbreviations*. London: Elsevier.

Johns, Ann M. and Tony Dudley-Evans. 1991. English for specific purposes: International in scope, specific in purpose. *TESOL Quarterly* 25/2: 297–314.

Kilgarriff, Adam. 2003. Linguistic search engine. In Simon Kiril ed. *Shallow Processing of Large Corpora: Workshop Held in Association with Corpus Linguistics*. Lancaster: University Centre for Computer Corpus Research on Language Technical Papers, 53–58.

Konstantakis, Nikolaos. 2007. Creating a business word list for teaching business English. *Estudios de Lingüística Inglesa Aplicada* 7: 79–102.

Lee, David and John Swales. 2006. A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes* 25/1: 56–75.

Long, Michael H. 1991. Focus on form: A design feature in language teaching methodology. *Foreign Language Research in Cross-cultural Perspective* 2/1: 39–52.

Malenica, France and Ivo Fabijanić. 2013. Abbreviations in English military terminology. *Brno Studies in English* 39/1: 59–87

Maniez, François. 2011. The contribution of corpus terminography specialized multilingual: The case of noun-adjective type name in the medical language. *Meta* 56/2: 391–406.

Navy Tactical Reference Publication NTRP-1-02. 2017. Washington DC: The Joint Staff.

Minkova, Donka. 2001. Review of the Cambridge History of the English Language: Vol. III—1476 to 1776. *Journal of English Linguistics* 29/1: 83–92.

Mishan, Freda. 2004. Authenticating corpora for language learning: A problem and its resolution. *ELT Journal* 58/3: 219–227.

Moon, Sungrim, Serguei Pakhomov, Nathan Liu, James O. Ryan and Genevieve Melton. 2013. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. *Journal of the American Medical Informatics Association* 21/2: 299–307.

Noguera-Díaz, Yolanda and Pascual Pérez-Paredes. 2019. Register analysis and ESP pedagogy: Noun-phrase modification in a corpus of English for military navy submariners. *English for Specific Purposes* 53: 118–130.

Plag, Ingo. 2003. *Word-formation in English*. Cambridge: Cambridge University Press.

Pérez-Paredes, Pascual. 2019. A systematic review of the uses and spread of corpora and Data-driven learning in CALL research during 2011–2015. *Computer Assisted Language Learning*: 1–26.

Pérez-Paredes, Pascual. 2020. *Corpus Linguistics for Education: A Guide for Research*. London: Routledge.

Pérez-Paredes, Pascual and Purificación Sánchez-Hernández. 2019. Uptake of corpus tools in the Spanish Higher Education context: A mixed-methods study. *Research in Corpus Linguistics* 6: 51–66.

Pérez-Paredes, Pascual, María Sánchez-Tornel, José Alcaraz Calero and Pilar Aguado-Jiménez. 2011. Tracking learners' actual uses of corpora: Guided vs. non-guided corpus consultation. *Computer Assisted Language Learning* 24/3: 233–253.

Pérez-Paredes, Pascual, María Sánchez-Tornel and José Alcaraz Calero. 2012. Learners' search patterns during corpus-based focus-on-form activities. *International Journal of Corpus Linguistics* 17: 483–516.

Plonsky, Luke and Frederick Oswald. 2014. How big is "big"? Interpreting effect sizes in L2 research. *Language Learning* 64/4: 878–912.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Spolsky, Bernard. 2004. *Language Policy*. Cambridge: Cambridge University Press.

Standard NATO Agreement 6001. 2019. *Language Proficiency Levels.* https://www.natobilc.org/en/products/stanags-60011142_stanag-6001

Stockwell, Robert and Donka Minkova. 2009. *English Words: History and Structure* (second edition). Cambridge: Cambridge University Press.

Taber, Keith. 2013. *Classroom-based Research and Evidence-based Practice: An Introduction*. London: Sage Publications.

The British Association for Applied Linguistics. 2006. *Recommendations on Good Practice in Applied linguistics*. http://www.baal.org.uk/dox/good

Thurston, Jennifer and Christopher N. Candlin. 1998. Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes* 17/3: 267–280.

Tono, Yukio, Yoshiho Satake and Aika Miura. 2014. The effects of using corpora on revision tasks in L2 writing with coded error feedback. *ReCALL* 26/2: 147–162.

Tribble, Chris and Glyn Jones. 1997. *Concordances in the Classroom: A Resource Guide for Teachers*. Houston: Athelstan.

Ushida, Eiko. 2005. The role of students' attitudes and motivation in second language learning in online language courses. *The Computer Assisted Language Instruction Consortium Journal* 23/1: 49–78.

Valipouri, Leila and Hossein Nassaji. 2013. A corpus-based study of academic vocabulary in chemistry research articles. *Journal of English for Academic Purposes* 12/4: 248–263.

Vyatkina, Nina. 2020. Corpus-informed pedagogy in a language course: Design, implementation, and evaluation. In Mariusz Kruk and Mark Peterson eds. *New Technological Applications for Foreign and Second Language Learning and Teaching*. Poland: IGI Global publishing company, 306–335.

Yoon, Hyunsook and Alan Hirvela. 2004. ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing* 13/4: 257–283.

Corresponding author
Yolanda Noguera-Díaz
Technical University of Cartagena
Department of Quantitative Methods, Legal Sciences and Modern Languages
Calle Real, 3.
30201. Cartagena.
Spain
e-mail: yolanda.noguera@upct.es

received: February 2020
accepted: June 2020

APPENDICES

APPENDIX 1A: Demographic information (10 informants; sailors)

1. Gender: 100% male

2. Mother tongue: 100% Spanish

3. Where did you last study English?
Secondary School: 70%
At University: 0%
At Military Schools: 30%
Others: 0%

4. What type of learning materials did you follow?
Books and workbooks: 80%
Blending learning (Books and online resources): 10%
OERs (Open Educational Resources): 10%

5. What is your performance level? Others?
CEFR: A1, A2, B1, B2, C1, C2.
NATO profiles:
0-70% (A1)
1-0%
2-20% (B2)
3-10% (C1)
4-0%

6. Have you studied general English or naval English in the Navy School?
Always general English: 80%
Always technical English related to the Navy: 0%
Fifty/fifty (general English and naval English): 10%
Sometimes naval English: 10%
Sometimes general English: 0%
Others: 0%

7. Some specific subjects were taught in English. If you remember the name of any, please, write it down.
Never: 100 %
Always: 0%
Often: 0%
Sometimes: 0%
Subject: …

8. Do you use a dictionary for writing tasks?
Yes: 80%
No: 20%
What type of dictionary?
Paper: 10%
On line: 90%
Others: 0%

9. Have you ever heard the term English for Specific Purposes?
Yes: 80%
No: 20%
Maybe: 0%
Reminder: English for Specific Purposes is related to particular disciplines. It has specific lexical, semantic and syntactic features of technical language. Its communicative functions convey their meaning in a unique way.

10. How often do you use a computer when studying English?
About once a day: 0%
About once a week: 40%
Never: 60%
Always: 0%
Others: 0%

11. How do you use the new technologies to improve your English skills? Choose the most suitable one.
a) Sometimes I use the Internet to look for grammar tutorials and similar: 80%
b) I often download podcasts and videos in English: 0%
c) I rarely use the new technologies, except for the CD player: 0%
d) I love surfing the net, reading, chatting or playing games with foreign people: 0%
e) I often use free/established digital didactic platforms to revise my English: 10%
f) I mainly use printed material: 0%
g) Other options: 10% (for music and chat)

12. Do you think the English Language is important for the Submarine crew in the Armed Forces?
a) If the Submarines are Spanish, the crew can speak Spanish to communicate the problems with the Base: 0%
b) The English language is only important when we sail in international waters in case of engine failures, damages or injured people: 50%
c) The International Submarine Escape and Rescue Liaison Office (Ismerlo) coordinates the rescue efforts from Norfolk: 10%
d) The Ismerlo protocols can also be translated into Spanish quickly: 0%
e) English language is only important for promoting: 0%
f) The high ranks must have a good English standard profile: 0%

APPENDIX 1B: Demographic information (6 informants; officers)

1. Gender: 80% male and 20% female

2. Mother tongue: 100% Spanish

3. Where did you last study English?
Secondary School: 100%
At University: 100%
At Military Schools: 100%
Others: 0%

4. What type of learning materials did you follow?
Classic: Book, workbook and media: 80%
Photocopies of different sources provided by the teacher and media: 0%
Blending learning: 10%
Open Educational Resources: 10%

5. What is your performance level? Others?
CEFR: A1, A2, B1, B2, C1, C2.
NATO profiles:
0-0% (A1)
1-0%
2-100% (B2)
3-0% (C1)
4-0%

6. Have you studied general English or naval English in the Navy School?
Always general English: 0%
Always technical English related to the Navy: 0%
Fifty/fifty (general English and naval English): 100%
Sometimes naval English: 0%
Sometimes general English: 0%
Others: 0%

7. Some specific subjects were taught in English. If you remember the name of any, please, write it down.
Never: 100 %
Always: 0%
Often: 0%
Sometimes: 0%
Subject: …

8. Do you use a dictionary for writing tasks?
Yes: 100%
No: 0%
What type of dictionary?
Paper: 0%
On line: 100%
Others: 0%

9. Have you ever heard the term English for Specific Purposes?
Yes: 80%
No: 20%
Maybe: 0%
Reminder: English for Specific Purposes is related to particular disciplines. It has specific lexical, semantic and syntactic features of technical language. Its communicative functions convey their meaning in a unique way.

10. Can you express the words *proa*, *popa*, *puente*, *escotilla* and *cabo* in English?
All of them: 100%
50%: 0%
I don´t remember now: 0%

11. With what type of content would you feel more comfortable in a role-play classroom activity, in a professional one or in a general one? Choose one.
a) Dialogue about the features of your current vessel: 0%
b) Dialogue about the Spanish/British weather: 0%
c) Dialogue about the protocols of safety on board: 100%
d) Dialogue about your spare time and hobbies: 0%

12. How often do you use a computer for studying English?
About once a day: 20%
About once a week: 0%
Never: 0%
Always: 80%
Others: 0%

13. How do you use the new technologies to improve your English skills? Choose the most suitable one.
a) Sometimes I use the Internet to look for grammar tutorials and similar: 80%
b) I often download podcasts and videos in English: 0%
c) I rarely use the new technologies, except for the CD player: 0%
d) I love surfing the net, reading, chatting or playing games with foreign people: 0%
e) I often use free/established digital didactic platforms to revise my English: 20%
f) I mainly use printed material: 0%
g) Other options: 0% (for music and chat)

14. Do you think the English language is very important for the submarine crew in the Armed Forces?
a) If the submarines are Spanish, the crew can speak Spanish to communicate the problems with the Base: 0%
b) The English language is only important when we sail in international waters in case of engine failures, damages or injured people. 0%
c) *The International Submarine Escape and Rescue Liaison Office* (Ismerlo) coordinates the rescue efforts from Norfolk: 0%
d) The Ismerlo protocols can also be translated into Spanish quickly 0%
e) English language is only important for promoting 0%
f) The high ranks must have a good English standard profile 70%
g) English language is the lingua franca for all sailors all over the world 30%

APPENDIX 2: Adapted from *The British Association for Applied Linguistics* (2006).

Academic Protocol: Interviews with students of the Navy Submarine School. Cartagena, Spain. May 2019.

Basis: Academic and Didactic Research Project on English Language Learning with Linguistic Corpora. Phase II.

Coordinator: Yolanda Noguera-Díaz. Lecturer at Technical University of Cartagena.

A) General responsibility with the informants (students):
-Anonymous and confidential identity (including gender and age). Numerical or alphabetical identification (e.g. student 1 or student A).
-Objectives and contents always of didactic type.
- Around 15 minutes of questions of a didactic nature in pairs or individually.
-Consent to record the answers with voice (zero image). Once the interviews between the researcher and the students have been transcribed, the audio files will be deleted.

B) Acceptance:
Once I have read the academic protocol and section A, I agree to participate in Phase II of this study in a totally anonymous and confidential manner.

In Cartagena, Spain, ... May 2019.
Signed: The informant

APPENDIX 3: Semi-structured interview questions

a) How do you find this approach at first sight?
b) How do you approach language learning on an everyday basis?
c) What is your goal as a language learner?
d) How do your balance your needs as an EFL learner and your needs as a military?
e) What is the role of vocabulary in your language learning?
f) How do you learn grammar and vocabulary?
g) Describe your experience with the DISSUB concordance lines.
h) How have concordance lines helped you understand and learn new language?
i) Have you found in concordance lines a good didactic method?
j) When are concordance lines useful to find out language patterns?
k) Can you focus our attention just on the right section of the line? Can you focus our attention just on the left section of the line?
l) Would you like to explore this Salvage and Rescue corpus with similar didactic exercises during a whole term?

RiCL Research in Corpus Linguistics

Building the *Great Recession News Corpus* (GRNC): A contemporary diachronic corpus of economy news in English

Javier Fernández-Cruz[ab] – Antonio Moreno-Ortiz[a]
Universidad de Málaga[a] / Spain
Pontificia Universidad Católica del Ecuador Sede Esmeraldas[b] / Ecuador

Abstract – The paper describes the process involved in developing the *Great Recession News Corpus* (GRNC), a specialized web corpus which contains a wide range of written texts obtained from the business section of *The Guardian* and *The New York Times* between 2007 and 2015. The corpus was compiled as the main resource in a sentiment analysis project on the economic/financial domain. A justification of the corpus design is provided, along with the methodology followed for the compilation process. To evaluate its usefulness, we include a sentiment analysis study on the evolution of the sentiment conveyed by the word *credit* during the years of the Great Recession.

Keywords – corpus linguistics; financial discourse; crisis studies; information retrieval; sentiment analysis

1. Introduction[1]

This paper describes and justifies the design and implementation of the *Great Recession News Corpus* (GRNC), a 21-million token compilation from 42,193 online business news articles of *The Guardian* and *The New York Times* published between 2007 and 2015. Our corpus serves as a useful linguistic resource for the scholarly research in multiple fields, such as English for Specific Purposes (ESP), comparative journalism or crisis studies, as it attempts to capture the impact that the Great Recession had on language.

The starting point of the GRNC is January 2007, coinciding with the emergence of the subprime crisis and the collapse of Lehman Brothers (2007–2008), which triggered a domino effect that was the foundation stone of the so-called 'credit crunch'. In addition, the corpus covers the Europe-centered aftershock, the 'sovereign debt crisis' in the

[1] This research was funded by the Pontifical Catholic University of Ecuador's Research Fund 2017: project LexiEcon. We would also like to thank the anonymous reviewers whose comments have greatly improved this manuscript.

Research in Corpus Linguistics 8: 28–45 (2020). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.08.02.02

European Union, which intensified from 2010 on and the impact of the “Whatever it takes” speech by Mario Draghi (European Central Bank 2012), and its final coverage coincides with the announcement of the European Central Bank’s *Quantitative Easing Program* (European Central Bank 2015). Apart from including the major economic events, the text compilation offers a full coverage of the socio-political crisis provoked by this instability, and the social response to a massive decline of the living standards in the daily lives of common people around the world.

Economic news coverage exposes the causes and reactions generated by this crisis. According to Lischinsky (2011: 154), different discursive frames “can have major effects in public understanding and policy decisions.” As a consequence, the GRNC allows to observe the discursive underpinnings of possible solutions from an ideological perspective at multiple linguistic levels (lexical, semantic, textual, etc.).

Another interesting factor here is the co-occurrence of the Great Recession with the decline of the so-called ‘old media’ or ‘legacy media’ (i.e., centralized printed newspapers or one-way broadcast technologies) and the rise of the age of digital media, which has been widely covered by the literature (Newman 2009; Huxford 2012; Franklin 2014). As the GRNC is composed entirely of web news, it may serve as witness to the innovation and radical changes across all aspects of a journalism already in search for alternative business models to start a sustainable journalism model for the future.

The origins of the GRNC go back to the research design of our main project: the development of a lexicon-based sentiment analysis (SA) system of financial texts with an appropriate treatment of the terminology in use during the Great Recession. In order to analyze these terms, a sentiment lexicon in the financial/economic domain, *SentiEcon*, was compiled from the corpus as a plugin lexicon for the *Lingmotif* sentiment analysis tool (Moreno-Ortiz 2017a, 2017b). Thus, corpus tools and techniques were used to (a) create a lexicon of sentiment words in the economic domain of the English language, and (b) to serve as a solid textual platform for observing the short-term diachronic —‘brachychronic’ if we follow Renouf’s (2002: 30) definition— evolution of sentiment in different lexical units within that domain.

In this paper we define the criteria used for selecting texts and we also explain the techniques we employed to process, organize, clean, and annotate the texts. For illustrative purposes, we also present a brief example of the research possibilities that this resource offers in sentiment analysis. Finally, we discuss its limitations and future perspectives.

2. JUSTIFICATION

The GRNC consists entirely of journalistic articles from the business section of the newspapers *The Guardian* and *The New York Times*; thus, the textual typology corresponds to a specialized corpus. Following the classification of publicly available corpora used by McEnery *et al*. (2006: 59), our corpus is a diachronic written monolingual corpus of business news. Since our aim was to study the evolution of sentiment conveyed by financial-economic terms as a result of the economic crisis, the GRNC is annotated by time of publication, covering a nine-year span (2007–2015), and organized monthly, resulting in a time series of 9x12 data points.

The analysis of the GRNC may provide an authentic overview of how news texts contribute to the linguistic construction of social reality. Both dailies publish with a view to influencing not only their readers, but also the discourse of the international press. In accordance with Bednarek and Caple (2012: 20–25), this motivation is justified, on the one hand, by the abundance of texts and, on the other, by the great exposure that the public has to news. Social reality is linguistically constructed, and such a construction is largely shaped by the view of journalists (Schudson 1989). When examining the use of crisis-related terms in leading media such as *The Guardian* or *The New York Times*, we are confronted with a type of language which narrates events to the public using carefully selected terms that depart from the specialized domain of economics. Both publications are highly authoritative internationally, and recognized for their stylistic influence and their ability to set the agenda (Van Belle 2003; Golan 2006).

The link between print media and the language of ordinary people is as old as the print itself, and reviews the central role of the popular press as a social educator (Conboy 2006: 9). The emergence of social media and the spread of viral news (Al-Rawi 2019) has served as an accelerator for the dissemination of new uses and meanings of words and terms through the articles that opinion makers have generated since it became widely available.

The novelty of this medium is bidirectionality. The online versions of traditional newspapers have progressively adapted to the needs of their online readers, who have become highly influential and, to a large extent, their discourse is modulated directly or indirectly by the mediation of its users through social media impact metrics or comments on social networks (Chung 2018). Nafría (2017: 236) argues that the challenge of adapting the headlines of *The New York Times* to Web 2.0 journalism implied a new discursive consolidation which required a combination of analytical journalism and simple language. As a result, newsrooms have designed social media policies to "guide newsworkers through the difficult intersection of traditional journalism and social media" (Duffy and Knight 2019: 932). Other relevant changes in online news exceed the textual level by incorporating multimodal items (e.g., animations or videos) or the appearance of new textual formats in their sections (i.e., blogs or microblogs). Paradoxically, the vast diversity of opinions that can be read on social networks (e.g., *Twitter*) has not diminished the influence of large media emporiums, but is thought to have increased the influence of traditional media on both the public and the stakeholders' opinion (for a thorough discussion, see Etter *et al*. 2017 and Blevins and Ragozzino 2019).

Due to the nature of our project, focused on sentiment analysis, news items are ideal for this task, as they are rich in evaluative language. Opinion is a key factor in the business sections of generalist media, since newswriters need to interpret macroeconomic figures and institutional statements in order to communicate this information to the public. Socioeconomic changes, as reflected in the texts, contribute to the construction of a value system, as understood by Thompson and Hunston (2000), which is built by the speaking community through evaluations. This system transcends as a component of ideology that permeates though the linguistic combinations and constructions of each of the texts. There is a vast array of definitions and discussions of the term 'evaluation' in linguistics and, in our view, Alba-Juez and Thompson's (2014: 13) is the most comprehensive one:

> a dynamical subsystem of language, permeating all linguistic levels and involving the expression of the speaker's or writer's attitude or stance towards, viewpoint on, or feelings about the entities or propositions that s/he is talking about, which entails relational work including the (possible and prototypically expected and subsequent) response of the hearer or (potential) audience. This relational work is generally related to the speaker's and/or the hearer's personal, group or cultural set of values.

The next step is to describe the features of the GRNC in order to implement a solid research framework. Corpora must be defined in terms of size, representativeness and balance (Xiao 2010: 148–153). As for size, Bowker and Pearson (2002: 49) consider that there is not a pre-established ideal number of words, since this depends mainly on the purpose of the study. While Sinclair's (1991: 18) maxim "a corpus should be as large as possible and should keep on growing" is still valid, even a small corpus can be a very useful resource if it is well designed. In particular, it is generally accepted that the size of a specialized corpus is generally smaller than that of a general corpus. Still, the GRNC is, however, significantly larger than other related corpora (see Section 3).

Representativeness is defined by Biber (1993: 243) as "the extent to which a sample includes the full range of variability in a population." Huan (2018: 57), however, questions this simplicity of operationalization, as different meanings of representativeness may emerge because, in contrast with general corpora, "most specialized corpora have already focused on special domain, time, and medium of the data." In our case, the main focus of the GRNC is hard news (domain) published online by two major British and American daily newspapers (medium) over the period between 2007 and 2015 (time).

The business section was selected because both newspapers fulfilled the following quality criteria: (1) the homogeneity of their language; (2) the editorial committees and the authors are representative experts of the domain; (3) an informative/didactic use of specialized language is made, so that it serves as a link between the specialist's discourse and the public; (4) the wide availability of the texts on the Internet; (5) the coverage of the main varieties of the English language; and (6) their online versions had free open access at the time of the compilation.

As for domain and medium representativeness, according to *ComScore* (2012), 644 million people worldwide accessed online newspaper sites in October 2012, representing 42.6% of the total Internet user base. Among reader popularity worldwide, *The New York Times* and *The Guardian* ranked second (48.7 million) and third (38.9 million), respectively. The business section of both dailies includes in-depth US/UK and international market news coverage, as well as company research tools. This section also includes international news involving political relations, finance and economy-related social issues. Texts include not only summaries of press conferences and economic reports, but also their interpretations, in the form of opinion columns, interviews, as well as live coverage of different events of interest and journalistic commentaries on the reactions of the public on social media. During the most turbulent events, both online sections included live coverage of major international events, such as meetings of the

Eurogroup. In addition to institutional coverage, possibly as a counterbalance, both media published crisis-related news related to the impact of the crisis on the common people, depicting social repercussions, such as the effects of mass unemployment, evictions, etc.

The aforementioned features do not qualify our corpus as representative, however. McEnery *et al*. (2006: 16) consider that specialized corpora are representative when the linguistic features at issue in the design are “subject to very limited variation beyond a certain point.” In relation to this, Huan (2018: 57) argues that the previous consideration of representativeness in specialized texts does not occur without criticism, since it involves examining the “linguistic variability” (lexical, syntactic, etc.) of a corpus at the expense of “situational variability” (i.e., the range of genres and registers in the target population). In any case, the linguistic criterion can serve to test the skewness of a corpus collected in line with the situational criterion. Finally, Tognini-Bonelli (2001: 57–59) considers that the representativeness of a corpus can hardly be evaluated in objective terms, and ultimately relates to the question of balance.

For Sinclair (2005: Section 5), balance implies that “the proportions of different kinds of text it [a corpus] contains should correspond with informed and intuitive judgements.” The balance of a corpus must be determined by the nature of the corpus and its intended research application (Xiao 2010: 149). McEnery *et al*. (2006: 16) debate the methodological problems behind Sinclair’s definition and contend that a reliable scientific measure of corpus balance has not been set. They also consider that any statement of corpus balance in the literature is very much an act of faith rather than a factual statement. In addition, Douglas (2003: 34) considers the balance of a corpus to be secondary to good research practice and, consequently, the resulting compilation must address research questions adequately and offer transparency in the documentation.

The GRNC also attempts to cover the two main varieties of English equally. Thus, the texts in *The Guardian* (British English) account for 47% of the corpus, while the remaining 53% was extracted from *The New York Times* (American English).

3. Related work

Corpora from the domains of economy, business and finance are compiled for diverse purposes (e.g., language for specific purposes, terminology or natural language processing). The growing awareness of Great Recession-related corpus research has led to different text compilations with a high disparity of sizes and purposes (i.e., discourse analysis, metaphor analysis, social network analysis, etc.). Nevertheless, to our knowledge, the GRNC fills an important gap, since no other English language corpus covers the main topics and features required for Great Recession journalistic or linguistic research. An array of examples of economics, business press and economic crisis-related corpora are reviewed synthetically in this section. In general terms, all prior work reviewed here can be included in one of these two genres: business communication or business news. The main business communication-related corpora are the following:

- The *Cambridge and Nottingham Spoken Business English Corpus* (CANBEC) (Handford 2010), one of the most widely distributed ESP corpora. It is an oral corpus that includes 912,734 words from 64 business meetings in 26 companies. It transcribes formal and informal meetings, presentations, phone conversations, etc.
- The *Hong Kong Financial Services Corpus* (HKFSC) (Li and Qian 2010) categorizes a total of 25 text types (among others, annual reports, fund description and speeches) in order to present a comprehensive picture of the written discourse in the financial services industry in Hong Kong. As of 2020, it is readily available online and includes more than 7 million words. It has been developed by the *Research Centre for Professional Communication in English* at the Department of English of the Hong Kong Polytechnic University.
- The *Malaysian Corpus of Financial English* (MaCFE) (Sadjirin *et al*. 2018) is a specialized online corpus which contains 4.3 million words from 1,472 electronic documents retrieved from banks and financial institutions’ official websites.
- Diesner *et al*. (2005) created a complex network corpus containing 252,000 corporate emails in order to observe the characteristics and patterns of communicative behavior of Enron employees during the different stages of its collapse.
- Lischinsky (2011) built a corpus of 50 financial and corporate social responsibility reports of Swedish companies in 2008 totaling 1.5 million tokens.

As for business news corpora, the following are noteworthy:

- The *Reuters Corpus Volume 1* (Rose *et al*. 2002) is a freely available archive of 806,791 English language *Reuters* news between 1996 and 1997. It covers news from different economic subdomains: corporate/industrial, economics, government/social and markets.

- Schröter and Storjohann (2015: 50) built a 4-million token “thematically homogenous ‘purpose-built’ corpus” which includes the keyword *financial crisis* in British newspaper articles from 2009.
- Rojo and Orts Llopis’ (2010) *English-Spanish Parallel Corpus* covers both *The Economist* and *El Economista* between two periods: the first one concerning the subprime crisis (June to November 2007), and the second one the era of the collapse of Lehman Brothers (September to December 2008).
- *Corpus de la Crisis Financiera* (CCF) (Botella *et al*. 2015) provide a snapshot of the opinion columns in Spanish daily papers *El País* and *El Mundo* throughout 2012.

4. Method for corpus compilation

In order to extract the texts, we decided to employ a custom semi-automatic procedure, since, despite the existence of many scrapers and other information extractors, no tools were found to fully satisfy our needs. We also intend to provide a concise and clear description of this pipeline in order to offer a simple, step-by-step guide for all levels of expertise. Our procedure may be summarized as follows:

1. Extraction of the URLs of each news item using mixed techniques.
2. Scraping of HTML files.
3. Extraction and cleaning of texts.
4. Classification, labelling and post-processing of corpus files.

In the first step, all the public URLs of the business section of both digital editions were extracted. To obtain good results, we used a monthly *Google Advanced Search*[2] to find all URLs containing the /business/ pattern from the domain of each newspaper.[3] All URLs were extracted with *Link Klipper* (2017), a simple yet very powerful browser extension, and then exported to a text file.

In order to scrape HTML files, we used the *Linux wget* tool, a simple command line utility for downloading files from the Internet. As input, we used text files containing the extracted hyperlinks, which allowed us to download all HTML files containing the news

[2] We are aware that *Google’s* personal search history can rearrange the order of matches. However, the influence of the said order is negligible since all the links were extracted.

[3] http://www.nytimes.com and http://www.guardian.co.uk.

items. Next, we used a custom shell script, available under demand, which classified the downloaded HTML files, both chronologically and by publisher, and discarded irrelevant (e.g., files containing no text) and repeated files. Finally, the files were cleaned automatically using the *BootCaT* utility (Baroni and Bernardini 2004) in order to keep labels such as <h1> or <p>, and discard all other irrelevant interface formatting elements. As a result, a typical corpus document contained headlines, sub-headlines and body text.

For an efficient search that allows us to observe the context of key terms chronologically, it was necessary to carry out a simple cataloguing of the texts. To this end, each of the files of the GRNC was named in a standard way to include the following coded metadata:

- The date of publication of the texts: encoded as four digits (YYMM, year-month). Thus, the date of a file published in August 2013 would be coded as “1308.”
- The name of the newspaper. In this case there are two *The Guardian* (GU) and *The New York Times* (NYT).
- A numeric ID code to identify each article, so that each of the text files received a unique identification code.

An example of a filename would be 1303NYT103.txt. This file would correspond to an article published in March 2013 on *The New York Times* with 103 as an identification number.

All text files were uploaded to *The Sketch Engine* (Kilgarriff *et al*. 2014) and subsequently compiled. As a result, all texts were tokenized and parsed automatically with *Penn Treebank POS-tagging Sketch Grammar for English TreeTagger version 3.1* (Marcus *et al*.1993).

5. CORPUS DESCRIPTION AND PRESENTATION

Table 1 summarizes the final composition of the GRNC: 42,193 texts containing 21.27 million words and 24.87 million tokens (i.e. words, punctuation, digit, abbreviations, product names and clitics). As for the lexicon, the corpus includes 242,000 different words grouped into 942,000 sentences.

Source	Tokens	Words	Texts	Sentences	%
The Guardian	13,197,301	11,285,112	21,312	477,165	53.06
The New York Times	11,673,704	9,982,273	20,881	465,753	46.93
TOTAL	24,871,005	21,267,385	42,193	942,918	100

Table 1: Description of the GRNC corpus by source

Figure 1 provides a breakdown of the number of tokens by year and publisher.

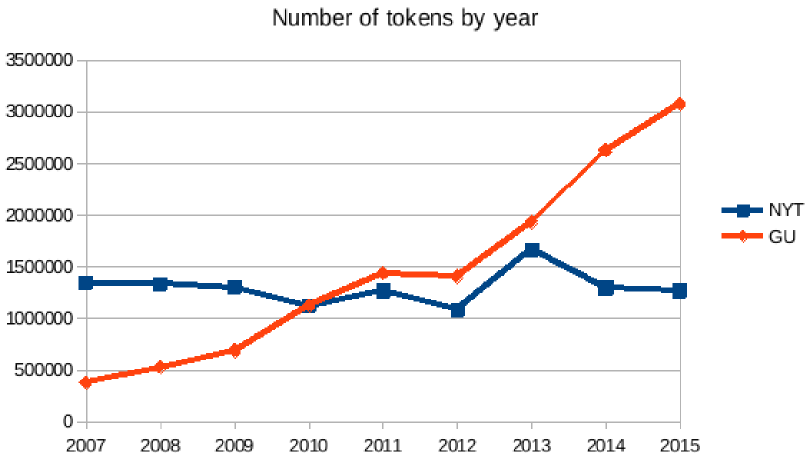


Figure 1: GRNC data number of tokens collected by year and publisher

The GRNC is available at *The Sketch Engine* by request for non-for-profit researchers. This platform was selected for its management, processing and dissemination possibilities, as well as the fact that our corpus can be used in combination with other 500 corpora in more than 90 languages that cover multiple language varieties.[4]

The GRNC can also be accessed as multiple subcorpora and, as a result, allow complex searches by year and publisher. For illustration purposes, a word frequency list from the corpus containing its most significant lexical items can be observed in Table 2.

[4] Visit http://tecnolengua.uma.es/grnc for more details. The corpus does not provide URLs as this data tend to change over time due to website rearrangements. For instance, *The New York Times* is currently behind a paywall.

Nouns		Adjectives		Verbs		Adverbs	
Lemma	**Freq.**	**Lemma**	**Freq.**	**Lemma**	**Freq.**	**Lemma**	**Freq.**
year	88,588	*more*	45,369	*be*	725,459	*not*	113,937
company	86,955	*new*	41,949	*have*	282,022	*also*	43,793
Mr.	57,976	*last*	39,677	*say*	182,133	*now*	27,440
business	55,435	*other*	33,712	*do*	71,287	*more*	27,437
market	38,291	*good*	23,764	*make*	52,053	*so*	24,910
people	35,747	*many*	23,179	*take*	35,327	*as*	22,641
bank	32,464	*big*	22,608	*go*	29,434	*just*	20,774
time	31,800	*chief*	21,624	*get*	27,609	*well*	19,668
price	29,283	*first*	20,579	*include*	27,217	*about*	19,171
sale	26,594	*high*	20,000	*use*	26,304	*even*	18,811
government	26,369	*financial*	19,682	*come*	24,870	*only*	17,179
percent	24,968	*large*	17,257	*work*	23,216	*still*	16,235
UK	24,502	*such*	16,084	*see*	21,385	*most*	13,463
month	24,187	*small*	13,654	*pay*	21,134	*very*	13,156
executive	23,704	*next*	13,219	*sell*	20,056	*then*	13,031
country	20,895	*economic*	12,472	*give*	19,138	*back*	11,939
share	20,586	*global*	12,139	*help*	18,766	*much*	11,294
industry	20,088	*low*	11,613	*need*	18,075	*too*	10,277
group	19,218	*own*	10,661	*want*	17,524	*already*	10,165

Table 2: Word frequency list in the GRNC

6. Sentiment analysis applications

We believe that this corpus offers a wide range of possibilities, from the observation of terms in use, or the analysis of new words or expressions in linguistics, to various applications in the digital humanities, such as contemporary historiography, and studies in behavioral economics, discourse analysis or compared media studies.

For illustration purposes, we briefly present here a study of the evolution of the sentiment conveyed by the term *credit*, an ‘event word’ (*mot événement*) during the Great Recession. According to Moirand (2007: 4), certain lexical units belong to a specific domain without connotations. After an event of a certain magnitude (e.g., the ongoing COVID19 crisis) that receives widespread media attention, these lexical units acquire connotative meanings related to this situation in particular.[5] As a consequence, these terms tend to appear in new contexts with new collocates that frequently may carry negative (or positive) sentiment and end up acquiring the sentiment of its collocates.

We extracted a data set of all the sentences from the corpus containing the keyword *credit* (*n*=6,764), and then proceeded to analyze it with the *Lingmotif* sentiment analysis software (Moreno-Ortiz 2017a, 2017b) in conjunction with the *SentiEcon* plugin lexicon

[5] Note the recent release of a brand new *Coronavirus Corpus*: https://www.english-corpora.org/corona/ (27 May, 2020.)

(Moreno-Ortiz *et al*. 2020). *SentiEcon* is a specialized sentiment lexicon on the financial domain. It contains 6,470 entries, both single and multi-word expressions, each with tags denoting their semantic orientation and intensity. It was extracted in its entirety from the GNRC. The main objective is to conduct a longitudinal study on the semantics of the word *credit*, from the sentiment perspective, by correlating the semantic orientation of the contexts in which this key term appears through the years that the corpus covers with the historical events that took place during that time.

In Figure 2 below, the resulting sentiment scores (first plot) and frequency trends (second plot) are presented in a time series. In order to smooth out random noise and seasonality in our plots, we calculated the yearly, rather than monthly, average of sentiment scores. We then used *The Sketch Engine* to extract the most frequent collocations of the term yearly. *LogDice* was selected as a statistic measure because it subsumes frequency and exclusivity of collocation. In addition, it is a standardized measure (range of 0–14) that avoids the bias produced by the different size of annual subcorpora (Gablasova *et al*. 2017: 164–166).

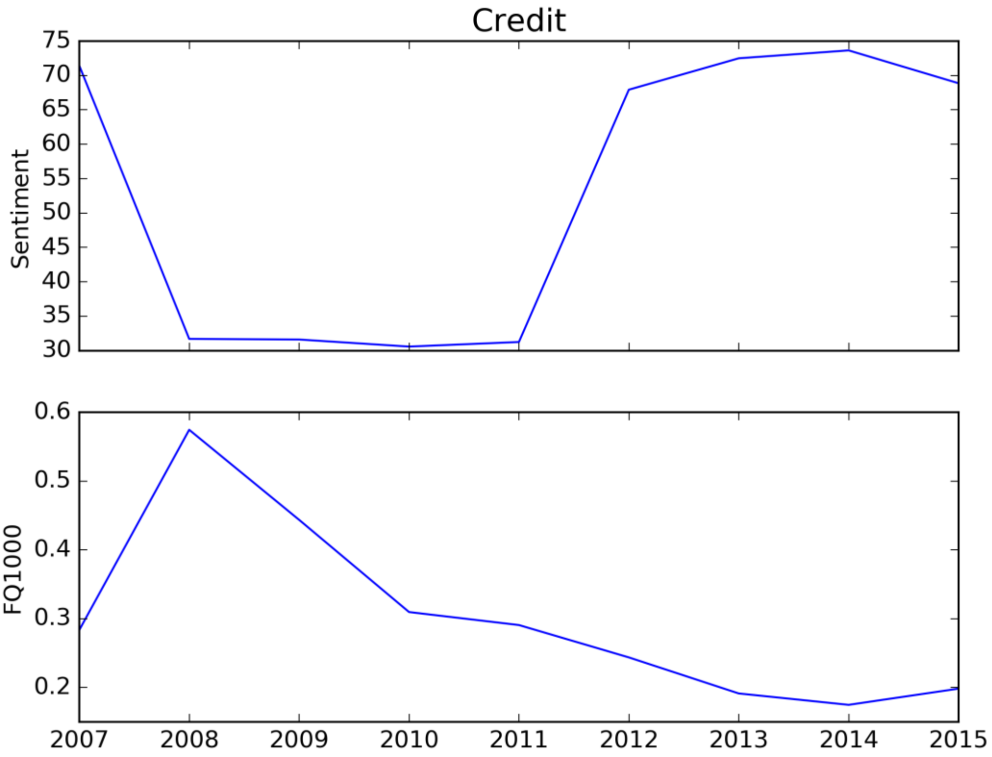


Figure 2: Sentiment and relative frequency plots for the term *credit*

In the sentiment plot, three clearly different trends can be observed:

a. The first stage corresponds to the dawn of the credit crisis in 2007, when *credit* was used in contexts with a positive semantic orientation. The analyzed sentences prioritized the use of specialized lexical items with neutral sentiment, including stable clusters (e.g., *credit card*, *credit line*) or other collocations with words such as *carbon*, *market* or *tax*, as illustrated in examples (1a) to (1d).

(1a) Gazprom, using Kyoto guidelines, plans to sell carbon CREDITS to Europe.

(1b) His business earns a tax CREDIT for hiring former prisoners.

(1c) Moreover, modern consumers love their CREDIT cards

(1d) Also, banks have traditionally had a monopoly on CREDIT and savings.

b. A second three-year phase (2008–2011) characterized by the sudden drop to a negative sentiment threshold. In addition, it can be observed that the relative frequency doubles in 2008 (0.57 per 1,000 words) compared to the previous year (0.28 per 1,000 words). These data correspond to the bursting of the housing bubble and the events that caused the credit system to freeze. The context of *credit* is characterized by more specific domain collocates that generally carry negative sentiment i.e., nouns: *default, squeeze* or *loss*, and adjectives such as *tight*, as illustrated in (2a) to (2d).

(2a) For a student, a default can destroy a CREDIT record, making it hard even to rent an apartment, let alone buy a home.

(2b) That simply doesn’t compare to the 150% bubbles we saw in some of the countries that were hit by the CREDIT crunch.

(2c) As the fund was being wound down, UBS said about 70 percent of its losses came from exposure to CREDIT default swaps.

(2d) Just as in the mortgage markets, a sterling CREDIT rating –the bond insurer’s seal of approval– is no longer trusted.

c. Semantic orientation is again reversed in 2012 and remains stable until 2015, while the relative frequency followed a slightly descending trend until the end of the series. It is pertinent to recall that by this time the journalistic machinery had set in motion a discourse in favor of reactivating credit from the central banks to the private banks. By then, the US Federal Reserve was consolidating its program of quantitative easing through the purchase of bank assets, and the new discursive paradigm following Mario Draghi’s famous speech in 2012 (European Central Bank 2012) was underway. Here, among the collocates of *credit*, we can find specific domain units and some previously absent positive items, e.g., *help, expand* or *cheap,* as illustrated in (3a) to (3d).

(3a) The SEC has disputed accusations that it has not done enough to tackle the individuals and companies that helped cause the credit crunch.

(3b) Legal to Censor, but Unwise Gabe Rottman, American Civil Liberties Union Pulling credit card services would help the haters and hurt free expression.

(3c) Cheap credit is essential when households and businesses are close to going bust.

(3d) The ECB has already taken steps to expand the supply of credit in an effort to drive down borrowing costs and ease pressure on household budgets.

It is then apparent that some level of correlation exists between the sentiment conveyed by the term *credit* and certain events that somehow determined its connotations. Of course, this simple study does not validate the corpus, but it certainly points to its usefulness as a research resource.

7. Conclusions, limitations and future perspectives

We have presented an ongoing project to design and build a diachronic, balanced, representative, and free-to-use corpus of economic-financial news from daily journals. Apart from our initial sentiment analysis application, the GRNC may be useful as a multipurpose resource, such as ESP, socio-economic studies, and diachronic linguistics.

Our corpus is still under development. Further research will shape the future of the GRNC, as our work is focused on the development of finely grained specific-domain sentiment analysis tools. One of the future goals is to expand its coverage to include (a) field-related texts from different journalistic sources and (b) non-journalistic sources, mainly social media and corporate reports.

The reason for expanding this corpus in these specific ways lies in the fact that the two newspapers which were used as sources share a similar liberal political angle. Future efforts will involve compiling other specialized publications of different ideologies, so that comparative language use can be performed. Another key factor in our future development is the question of the study of the expression of economic language from different levels of specialization.

On the other hand, integrating social media sources would allow us to compare the use that the public makes of economic language. Sources such as blogs, online comments in newspapers and social media would undoubtedly enhance the possibilities of the current corpus. Other potential research possibilities would involve comparative studies on terminological trends in order to determine the level of influence of institutions and mainstream media into the general public.

Finally, the observation of highly specialized language from documents issued for specialists is a field of special interest, as is the case of internal corporate disclosures. In this way, the lexical, cognitive and affective divergences between different levels of specialization could be observed: specialist discourse, journalistic/informative language and public use of specialized terms.

References

Al-Rawi, Ahmed. 2019. Viral news on social media. *Digital Journalism* 7/1: 63–79.

Alba-Juez, Laura and Geoff Thompson. 2014. The many faces and phases of evaluation. In Laura Alba-Juez and Geoff Thompson eds. *Evaluation in Context.* Amsterdam: John Benjamins, 3–24.

Baroni, Marco and Bernardini, Silvia. 2004. *BootCaT*: Bootstrapping corpora and terms from the web. In María Tersa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa and Raquel Silva eds. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Paris: European Language Resources Association, 1313–1316.

Bednarek, Monika and Helen Caple. 2012. *News Discourse*. London: A&C Black.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8/4: 243–257.

Blevins, Dane P. and Roberto Ragozzino. 2019. On social media and the formation of organizational reputation: How social media are increasing cohesion between organizational reputation and traditional media for stakeholders. *Academy of Management Review* 44/1: 219–222.

Botella, Ana, Keith Stuart and Lucía Gadea. 2015. A journalistic corpus: A methodology for the analysis of the financial crisis in Spain. *Procedia – Social and Behavioral Sciences* 198: 42–51.

Bowker, Lynne and Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge.

Chung, Jae Eun. 2018. Peer influence of online comments in newspapers: Applying social norms and the social identification model of deindividuation effects (SIDE): *Social Science Computer Review* 36/5: 551–567.

ComScore. 2012. *Most Read Online Newspapers in the World: Mail Online, New York Times and The Guardian*. https://www.comscore.com/Insights/Infographics/Most-Read-Online-Newspapers-in-the-World-Mail-Online-New-York-Times-and-The-Guardian (4 May, 2020.)

Conboy, Martin. 2006. *Tabloid Britain: Constructing a Community through Language*. London: Routledge.

Diesner, Jana, Terril L. Frantz and Kathleen M. Carley. 2005. Communication networks from the *Enron Email Corpus* "It's always about the people. Enron is no Different." *Computational and Mathematical Organization Theory* 11/3: 201–228.

Douglas, Fiona M. 2003. *The Scottish Corpus of Texts and Speech*: Problems of corpus design. *Literary and Linguistic Computing* 18/1: 23–37.

Duffy, Andrew and Megan Knight. 2019. Don't be stupid. *Journalism Studies* 20/7: 932–951.

Etter, Michael, Davide Ravasi and Elanor Colleoni. 2017. Social media and the formation of organizational reputation. *Academy of Management Review* 44/1: 28–52.

European Central Bank. 2012. Verbatim of the remarks made by Mario Draghi. Speech given at the Global Investment Conference. London, 26 July 2012. https://www.ecb.europa.eu/press/key/date/2012/html/sp120726.en.html (25 May, 2020.)

European Central Bank. 2015. Introductory statement to the press conference (with Q&A) by Mario Draghi. Frankfurt am Main, 22 January 2015. https://www.ecb.europa.eu/press/pressconf/2015/html/is150122.en.html (25 May, 2020.)

Franklin, Bob. 2014. The future of journalism. *Journalism Studies* 15/5: 481–499.

Gablasova, Dana, Vaclav Brezina and Tony McEnery. 2017. Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning* 67/1: 155–179.

Golan, Guy. 2006. Inter-media agenda setting and global news coverage. *Journalism Studies* 7/2: 323–333.

Handford, Michael. 2010. *The Language of Business Meetings*. Cambridge: Cambridge University Press.

Huan, Changpeng. 2018. *Journalistic Stance in Chinese and Australian Hard News*. Shanghai: Springer.

Huxford, John. 2012. Reporting on recession: Journalism, prediction, and the economy. *International Business & Economics Research Journal (IBER)* 11/3: 343–356.

Kilgarriff, Adam, Vit Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel. 2014. *The Sketch Engine*: Ten years on. *Lexicography* 1/1: 7–36.

Li, Yongyan and David D. Qian. 2010. Profiling the academic word list (AWL) in a financial corpus. *System* 38/3: 402–411.

Link Klipper 1.0.0. 2017. http://www.codebox.in/products/linkklipper/ (7 May, 2020.)

Lischinsky, Alon. 2011. In times of crisis: A corpus approach to the construction of the global financial crisis in annual reports. *Critical Discourse Studies* 8/3: 153–168.

Marcus, Mitchell P., Beatrice Santorini and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19/2: 313–330.

McEnery, Tony, Richard Xiao and Yukio Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.

Moirand, Sophie. 2007. *Les Discours de la Presse Quotidienne. Observer, Analyser, Comprendre*. Paris: Presses Universitaires de France.

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

Moreno-Ortiz, Antonio. 2017b. Lingmotif: Sentiment analysis for the digital humanities. In Mirella Lapata, Phil Blunsom and Alexander Koller eds. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia: Association for Computational Linguistics, 73–76.

Moreno-Ortiz, Antonio, Javier Fernández-Cruz and Chantal Pérez-Hernández. 2020. Design and evaluation of SentiEcon: A fine-grained economic/financial sentiment lexicon from a corpus of business news. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asución Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*. Marseille: European Language Resources Association, 5067–5074.

Nafría, Ismael. 2017. *La Reinvención del New York Times: Cómo la Dama Gris del Periodismo se está Adaptando.* Austin: Knight Center.

Newman, Nic. 2009. *The Rise of Social Media and its Impact on Mainstream Journalism*. Oxford: Reuters Institute for the Study of Journalism, Department of Politics and International Relations, University of Oxford.

Renouf, Antoinette. 2002. The time dimension in modern English corpus linguistics. In Bernhard Kettemann and Georg Marko eds. *Teaching and Learning by Doing Corpus Analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19-24 July, 2000*. Amsterdam: Brill/Rodopi, 27–41.

Rojo López, Ana María and María Ángeles Orts Llopis. 2010. Metaphorical pattern analysis in financial texts: Framing the crisis in positive or negative metaphorical terms. *Journal of Pragmatics* 42/12: 3300–3313.

Rose, Tony, Mark Stevenson and Miles Whitehead. 2002. *The Reuters Corpus Volume 1*– From yesterday's news to tomorrow's language resources. In Manuel González Rodríguez and Carmen Paz Suárez Araujo eds. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. Las Palmas de Gran Canaria: European Language Resources Association, 827–833.

Sadjirin, Roslan, Roslina Aziz, Nordin Abdul, Ismail Mohd Rozaidi and Norzie Diana Baharum. 2018. The development of *Malaysian Corpus of Financial English* (MaCFE). *Journal of Language Studies* 18/3: 73–100.

Schröter, Melani and Petra Storjohann. 2015. Patterns of discourse semantics: A corpus-assisted study of financial crisis in British newspaper discourse in 2009. *Pragmatics and Society* 6/1: 43–66.

Schudson, Michael. 1989. The sociology of news production. *Media, Culture & Society* 11/3: 263–282.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John. 2005. Corpus and text – Basic principles. In Martin Wynne ed. *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books. http://users.ox.ac.uk/~martinw/dlc/index.htm (7 May, 2020.)

Thompson, Geoff and Susan Hunston. 2000. Evaluation: An introduction. In Susan Hunston and Geoff Thompson eds. *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: Oxford University Press, 1–26.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work.* Amsterdam: John Benjamins.

Van Belle, Douglas A. 2003. Bureaucratic responsiveness to the news media: Comparing the influence of *The New York Times* and network television news coverage on US foreign aid allocations. *Political Communication* 20/3: 263–285.

Xiao, Richard. 2010. Corpus creation. In Nitin Indurkhya and Frederick J. Damerau eds. *Handbook of Natural Language Processing.* Boca Raton: Chapman & Hall/CRC, 147–165.

Corresponding author:
Javier Fernández-Cruz
Departamento de Filología Inglesa, Francesa y Alemana
Escuela de Ingenierías Industriales
C/ Doctor Ortiz Ramos, s/n
29071 Málaga, Spain
fernandezcruz@uma.es

received: February 2020
accepted: June 2020

RiCL Research in Corpus Linguistics

Building a parallel corpus of literary texts featuring onomatopoeias: ONPACOR

Aroa Orrequia-Barea[a] – Cristian Marín-Honor[b]
University of Jaén[a] / Spain
University of Cádiz[b] / Spain

Abstract – Onomatopoeias constitute a much neglected subject in linguistics. The rather scarce literature on onomatopoeias is derived from a lack of reliable empirical data on the topic. In order to bridge this gap, we have compiled a parallel corpus of literary texts featuring onomatopoeias: the *Onomatopoeia Parallel Corpus* (ONPACOR). The corpus consists of onomatopoeias in English, Spanish and French extracted from comics and representative corpora of each language. ONPACOR has been built on the basis of existing translations to the languages of reference. This article describes the methodology used to compile the corpus, as well as the applications that it can have.

Keywords – corpus linguistics; lexicology; onomatopoeias; ONPACOR; parallel corpus; web application

1. INTRODUCTION[1]

The fact that onomatopoeias have not been thoroughly studied in linguistics is probably due to the lack of agreement regarding their categorial status (Barbéris 1992: 52). Admittedly, there have been several attempts to compile dictionaries of onomatopoeias such as Gasca and Gubern (2008), a dictionary of onomatopoeias used in Spanish comics, or Enckell and Rézeau (2003), a dictionary of onomatopoeias in French. There have also been attempts to include onomatopoeic forms in general dictionaries such as the *Diccionario de la Lengua Española* (2020) in Spanish or the *Merriam Webster*

[1] We wish to express our gratitude to María Luisa Mora Millán (University of Cádiz) and José María García Núñez (University of Cádiz) for providing insight and assistance during this project. We must also thank the anonymous reviewers and the editors for their comments, which have greatly improved the manuscript. Finally, we are also immensely grateful to Encarnación Almazán Ruiz (University of Jaén) and Alejandro Alcaraz Sintes (University of Jaén) for their invaluable help during the revision process.

Research in Corpus Linguistics 8/2: 46–62 (2020). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.08.02.03

Dictionary in English. However, these collections are not very thorough, since these dictionaries are not specifically devoted to onomatopoeias.

This lack of multilingual resources on onomatopoeias has led us to compile the *Onomatopoeia Parallel Corpus* (ONPACOR). This parallel corpus consists of a set of onomatopoeic words in English, Spanish and French retrieved from representative corpora in each language, namely, *The British National Corpus* (BNC) for English*,* the *Corpus de Referencia del Español Actual* (CREA; Real Academia Española 2020) for Spanish and the *FRANTEXT* corpus (ATILF 2019) for French. ONPACOR has been built on the basis of existing translations of the three languages of reference. For this reason we have restricted our search to literary texts.

The article is organised as follows. Section 2 offers a review of the literature, establishes the distinction between interjections and onomatopoeic expressions, and dwells on the translation problems onomatopoeias pose. Section 3 describes the methodology used to build our parallel corpus. In Section 4, we mention some applications and further developments of the corpus. Section 5 provides information on further research. Finally, Section 6 offers a summary and some conclusions.

2. Review of the literature

As already mentioned, there has been little linguistic research on onomatopoeias. There are several reasons for this. On the one hand, there is no real consensus as regards their definition. Additionally, the grammatical status of onomatopoeias is uncertain, which leads to an incorrect consideration of onomatopoeias and interjections as belonging to the same class (Melnikienė 2016: 169). On the other hand, although some authors, such as de Saussure (2011: 69), claim that the number of onomatopoeic words in languages is very small, in actual fact the prevalence of onomatopoeias is language-specific.

Before launching into the corpus compilation process, it seemed essential to define both onomatopoeias and interjections as clearly as possible. According to the *Merriam Webster Dictionary*, an onomatopoeia is “the naming of a thing or action by vocal imitation of the sound associated with it (such as *buzz, hiss*).” In other words, they are part of the “creative process of the conventional language” (Arboleda and Arce-Lopera 2017: 172) in order to reproduce human, animal, natural or artificial sounds (de la Rosa Regot 2015: 2), that is, an onomatopoeia constitutes “an imitative-driven transformation of a sound of nature into a word" (Assaneo *et al.* 2011). According to de Saussure (2011: 69), onomatopoeic words are one of the exceptions to the so-called arbitrariness of the linguistic sign, insofar as their phonological forms are not arbitrary, but clearly seem to be linked to their meaning. For these reasons, some authors (Sugahara 2011: 2) do not consider them to be, strictly speaking, parts of speech.

The fact that onomatopoeias are imitations of sounds may lead to the conclusion that their representation in language does not vary crosslinguistically. Thus, "one could ideally expect that the imitation of a simple noise should be a single speech sound, the closest one from an acoustical point of view" (Assaneo *et al.* 2011), since sounds are identical regardless of the language system. However, contrary to this intuition, every language has its own onomatopoeic words made up of the consonants and vowels belonging to the phonological system of each language, with clearly different and distinctive properties. As a consequence, onomatopoeias are usually assimilated to the phonological system of that language, causing variations in the shape of words from one language to another in their attempt to represent the same sound (Assaneo *et al.* 2011). An obvious, and much cited, example of this is the onomatopoeic word representing the rooster's crow in different languages. In Spanish it is *quiquiriqui*, but in French it is *cocorico* and in English *cock-a-doodle-do* (de Buron-Brun 2006: 768).

Nevertheless, the imitation can vary from an almost complete match between the sound and the word representing it —where the human vocal possibilities are exploited so as to obtain the closest imitation to a non-human sound, as is the case in animal onomatopoeias— to a more imprecise representation, in which the mapping between the sound and the word is just an approximation (Rhodes 2010: 279). This approximate imitation supports the theory of 'Sound Symbolism', which posits that speakers invest certain phonemic sounds with particular meanings. Therefore, certain phonemes seem to embody certain non-acoustic properties, such as the use of, for example, the front closed high vowel *i*, which conveys the idea of a non-human sharp high-pitched noise, as in *clink* or *click*,[2] while at the same time it seems to trigger the idea of a diminutive sense (Rhodes 2010: 284). Rhodes (2010: 280), in fact, states that structured sound symbolism "deals with regularities of relationships between forms and meaning that do not readily appear in traditional morphological analysis."

[2] In the *Merriam Webster Dictionary*, *click* is defined as "a slight sharp noise" and *clink* as "a slight sharp short metallic sound."

Even though onomatopoeias in all languages are an extremely tiny set of words as compared to other classes, e.g. verbs, there seem to exist languages —such as Japanese— which are more onomatopoeia-prone than others, such as English or Spanish (Sugahara 2011: 1). In any case, onomatopoeias are frequently used in people's daily life to "complement and enrich their verbal or written communication" (Arboleda and Arce-Lopera 2017: 172) and to express feelings and emotions, which cannot always be expounded by using just words. In fact, Arboleda and Arce-Lopera (2017: 172) studied the widespread use of onomatopoeic words to express speakers' experience with food and concluded that "compared to adjectives (e.g., *soft*, *smooth*), onomatopoeias can better describe a texture because they use symbolic sounds to represent a sensory experience."

As far as interjections are concerned, these are words related to the expression of feelings (Melnikienė 2016: 177). According to the *Merriam Webster Dictionary*, an interjection is "an ejaculatory utterance usually lacking grammatical connection: such as (a) a word or phrase used in exclamations (such as *Heavens! Dear me!*); (b) a cry or inarticulate utterance (such as *Alas! Ouch! Phooey! Ugh!*) expressing an emotion." Melnikienė (2016), drawing on Barbéris (1992) and Swiatkowska (2000), distinguishes between two different types of interjections: 'modal' and 'dictal'. On the one hand, modal interjections refer to fixed words belonging to any word class which express emotions, attitudes or feelings (Melnikienė 2016: 177). For example, the Spanish interjection *ay* can be used either in a context of surprise or in a context of pain. As Barbéris (1992: 52) states, the distinguishing feature of this type of interjections is that they do not reflect any aspect of the world, but the emotions of the subject (Melnikienė 2016: 177). On the other hand, dictal interjections refer to the imitation or adaptation of natural sounds and, as a result, this type of words are impersonal and do not express emotions or feelings (Melnikienė 2016: 178).

The main problem to distinguish between onomatopoeias and interjections lies in that most onomatopoeias are considered interjections. However, there are many interjections that do not have an onomatopoeic origin (Kleiber 2006: 11). This can lead to confuse them, basically because some onomatopoeias can be used as interjections. A clear example is the Spanish onomatopoeic expression of laughter *ja*, which is also used as an interjection to show disagreement or disbelief (Husillos Ruiz 2018: 18). The main difficulty arises with onomatopoeias that imitate human sounds, which is where the boundaries between both words become blurred. This is the reason why these two types of concepts seem to intersect in some contexts (Kleiber 2006: 10).

Despite the aforementioned problem, there are clear differences between interjections and onomatopoeias. First, interjections are considered an independent word class, a part of speech, whereas onomatopoeias are not. Both of them are invariable words and independent as far as their syntactic properties are concerned, but only interjections are able to express meaningful speech acts. In fact, interjections can be the head of syntactic groups or even make up locutions, whereas onomatopoeias cannot. For Kleiber (2006: 12), interjections are "phrases with an implicit predication," which means that they have a semantic import. For instance, the French modal interjection *pouah!* is used by speakers to convey their strong repugnance at something disgusting (Kleiber 2006: 12).

Regarding morphology, apart from the word class of interjections proper, other word classes, such as nouns or adjectives, can also do duty as interjections. For example, in Spanish, the noun *caracoles* can be used as an interjection to show anger, surprise or admiration (Husillos Ruiz 2018: 21). Nevertheless, all these words have something in common: they are used to communicate the speaker's natural sounds of feelings and emotions (de Buron-Brun 2006: 768).

From a pragmatic and semantic point of view, onomatopoeias tend to lack semantic content because they are restricted to the imitation of sounds. However, they are influenced by extralinguistic factors or the context in which they are used, sometimes even in visual material, such as comics. For this reason, it is very difficult to associate each onomatopoeia to just one meaning. In most cases, one onomatopoeia may represent more than one sound. To give a simple example, the Spanish word *pum* can refer to either a gunshot or a door knock (Orrequia-Barea and Marín-Honor 2018: 97). Certainly, attempts have been made to tackle this problem by means of lists of onomatopoeias or glossaries, though most often lacking reliability (Sugahara 2011: 2). Besides, certain scholars, such as Sugahara (2011), have already focused on the onomatopoeia translation problem. Additionally, de la Rosa Regot (2015) provides strategies for translating onomatopoeias between English and Spanish. Similarly, Husillos Ruiz (2018) compiled a multilingual glossary and conducted research in translation strategies in advertising. Also, there have been attempts to compile dictionaries of onomatopoeias, such as Gasca and Gubern (2008) in Spanish, and

Enckell and Rézeau (2003) in French. Despite this body of research, little consideration has been given to crosslinguistic differences, and there is a dramatic lack of multilingual resources. The compilation of a multilingual corpus or even a multilingual dictionary focused on onomatopoeic words may help solve translation-related problems among languages. Interestingly, onomatopoeias are not just words which imitate sounds, but they are also used to help language users complement and enhance their communication. For this reason, it is necessary to translate them because not doing so may cause a loss of meaning. Since one single non-human sound may be conveyed by means of different onomatopoeic expressions, it seems necessary to have a multilingual resource containing all these equivalent words or expressions. We believe that the *Onomatopoeia Parallel Corpus* (ONPACOR) constitutes a major step in that direction.

3. Methodology

In this section we discuss the process of compilation of ONPACOR. We have followed four main steps: the compilation of a list of onomatopoeias, the search for the concordances, the translation phase and the creation of the interface. These steps will be explained in what follows.

3.1. Compilation of onomatopoeias

Two different strategies were implemented to make the compilation based on the availability of reference corpora. Our first idea was to extract onomatopoeias from corpora of each language since we wanted to have empirical evidence that those onomatopoeic forms were actually used in the language. For this reason, we intended to download the corpora to look for onomatopoeias using regular expressions to get as many onomatopoeic forms as possible without restricting them to the most common ones. However, we could only follow this procedure with the BNC, since it was the only corpus that could be downloaded. For Spanish and French, the CREA and *FRANTEXT* corpora were not downloadable, so that we had to follow a different process, namely manually extracting onomatopoeias from comics. Nonetheless, we extracted as many instances from comics as we found in the BNC so that the sample would be balanced.

For Spanish, we chose CREA, which includes oral and written texts produced in every Spanish-speaking country from 1975 until 2004. The main reason for the selection of this corpus was that 90% of the texts are written, whereas only 10% represent the spoken language. As for French, the corpus chosen is *FRANTEXT*, which basically contains literary and philosophical texts from 1180 to 2009, which suited the purpose of our corpus. Finally, regarding English, we worked with the BNC, of which 90% belong to written texts from the late twentieth century. These three corpora were suitable for the purpose of our project, since they consist of written texts, which made it easier to find the translations.

3.1.1. English

Firstly, we made an analysis of the form of the onomatopoeic words in the BNC in order to compile them. The purpose of this analysis was to find patterns in the formation of onomatopoeias. Therefore, we systematised some typical combinations found in onomatopoeias in three groups: 1) vowels, 2) consonants and 3) endings of words (Kwon 2015: 39–71).[3]

1. Long vowels tend to be used to represent slow movements while short tend to represent quick ones. Additionally, onomatopoeias often consist of similar words with a change of vowels to represent two-phased movements, such as *ding-dong* or *flip-flop*. Vowels in ablaut-like alternation are usually repeated twice or more, sometimes even three or four times.
2. Consonants are usually found in pairs to represent different movements and ways of doing things. Most typically we find the combinations: consonant + *-l* or *-r*.
3. Endings of words are usually made up of two-consonant clusters or repeated vowels.

All the above-mentioned systematisations were captured by means of regular expressions, which are patterns that are frequently used in text editors to look for,

[3] Kwon's (2015: 40) research is focused on 'phonaesthemes', that is, "recurrent pairings of sound and meaning." Although we do not deal with phonaesthemes here, Kwon's study is used as a basis to systematise some patterns of sound symbolism that are expected to be found in onomatopoeias.

substitute and replace a sequence of characters. This sequence has to fulfil the criteria set out by the regular expression.[4]

As the main purpose was to find most of the onomatopoeias in the BNC, the following regular expressions, based on the previous patterns of formation, were used:

1. To find consonants that were repeated at least three times: [bc-df-hj-np-tv-z]{3}. This regular expression yielded onomatopoeias such as *zzz*.
2. To find the pattern of up to two consonants plus vowels, repeated at least twice, followed optionally by an indefinite number of consonants: [bc-df-hj-np-tv-z]{0,2}vowel{2,} [bc-df-hj-np-tv-z]{0,}. We typed each of the five different vowel graphemes in the vowel slot. Some of the results were: *craark*, *beep*, *riing*, *boom* or *uuummm*.

3.1.2. Spanish and French

Comic books are known to be a very rich repository of onomatopoeias. That is the reason why we manually compiled a list of onomatopoeic words from the most popular comics in each language. For Spanish, we chose a variety of issues from some of the most popular comics in Spain, among others, *Zipi y Zape* by José Escobar, *Mortadelo y Filemón* and *Rompetechos* by Francisco Ibáñez, *El doctor Cataplasma* by Martz Schmidt and *Sir Tim O'Theo* by Raf. As for French, we chose such popular comics as *Astérix et Obélix* by René Goscinny and Albert Uderzo, *Spirou et Fantasio* by Rob-Vel and *Les Aventures de Tintin et Milou* by Georges Remi. As a result, we achieved a compilation of 500 onomatopoeias.

3.2. Concordances

After compiling the three lists of onomatopoeias, the next step was to check whether there were concordances with those onomatopoeias in each corpus. As a matter of fact, the motivation to select each corpus was based on the existence of written literary texts, so that we could find the translations in the following step. Although onomatopoeias are frequently used in dialogues and in the spoken language, it was not until the nineteenth

[4] For further information on regular expressions see https://regexr.com/.

century and the birth of Realism that onomatopoeias started to appear in novels, in an attempt by writers to reproduce the colloquial language (Bueno Pérez 1994: 15).

The process of looking for concordances was basically the same for the three languages. We searched for each onomatopoeia in the chosen corpus and then stored all its concordances. We also stored some interesting metadata, which was included in each concordance, such as the year of publication, the author and the title of the work. In the case of the CREA interface, there is a book filter which makes it easier to find the translation. Since we downloaded the BNC, we used *AntConc* (Anthony 2019) for the retrieval of the data.

3.3. Translations

In order to be able to find existing translations of the texts, at this stage we needed to use corpora mainly containing written or literary texts. For this purpose, we used the website *Index Translationum*, a database of book translations promoted by UNESCO.[5] The user can set up the search criteria for any given book, so that the database provides a record of the translations in all different languages. In order to restrict the search, we provided just the author's name and the title. As previously mentioned, this information was obtained in the second step. The list of records contains such information about each translation as the title of the work in the target version, the language used in the translation, the translator, the place and the year of publication, among others. This is illustrated in (1) and (2), which show a search for *The Colour of Magic* by Terry Pratchett, which returned 36 hits.

(1) Pratchett, Terry: *El color de la magia* [Spanish] / Macía, Cristina / México, D.F.: Roca [México], 1989. 224 p. The Colour of Magic [English]

(2) Pratchett, Terry: *La huitième couleur* [French] / Marcel, Patrick / Nantes: l'Atalante [France], 1993. 283 p. The Colour of Magic [English]

This database was used to filter out the titles which lacked translations in the other languages. Once the translation of the book was found, the exact excerpt —namely the section which included the onomatopoeia— was extracted.

[5] http://www.unesco.org/xtrans/.

3.4. Implementing the interface

ONPACOR will be hosted on an online platform so that users can access it easily.[6] This platform is actually a web application, which can be operated in a browser. To implement this platform, we needed to design a database to host the texts featuring the onomatopoeias in the three different languages. For the design of the database, we used a conceptual model (Coronel and Morris 2016: 71–274), because it represents a comprehensive picture of the information, as the users are going to see it, ignoring implementation details as well as the structure of the information, which makes it more understandable for researchers in the Humanities. To shape this information, we used an 'entity-relationship model' (Chen 1976), which is the most widely used conceptual model. It is made up of a set of concepts which allows to describe reality by means of a set of linguistic and graphic representations, presenting a natural vision of the real world. Moreover, this model has a number of advantages. First, it only reflects the existence of the information. Secondly, it does not depend on any particular database or operating system. Thirdly, it is open thus allowing the system to be updated as much as possible, which is a great advantage for an on-going project such as this one. The major steps taken to create the database are presented in the following paragraphs.

The first step was to generate a universe of discourse (Boole 1854), namely, a description of the information, the collection of objects that will be included in the database as well as a schema about how these data are related. In ONPACOR, the universe of discourse is made up of the onomatopoeias, which were extracted in the compilation step. These onomatopoeias are related to the concordances that were extracted from the corpus, and, at the same time, they are related to the translations in the other two languages.

The second step was to build the database, for which purpose we used a standard database management system, *MySQL,* for two main reasons: on the one hand, it is one of the most commonly used open source databases and, on the other hand, it is used for relational databases, which fits perfectly with ours. To build the database, we used two tables, one for the concordances and another for the onomatopoeia. As can be seen in Figure 1, some information is required in the concordance table, namely, the language, the concordance itself and each translation. Likewise, it is possible to highlight whether

[6] A similar platform has already been implemented to host the online dictionary of onomatopoeias in Spanish (Orrequia-Barea and Marín-Honor 2018).

one particular extract belongs to the original version. The information that is going to be introduced into this database may be restricted by using the parameters found in *MySQL*. This includes a 'Variable chain of characters' (varchar) or a 'Boolean operator' (bool) to indicate whether the excerpt is original or not, as shown in Figure 1.

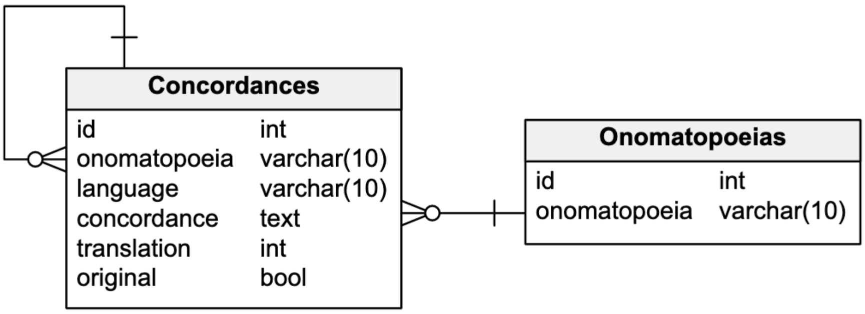
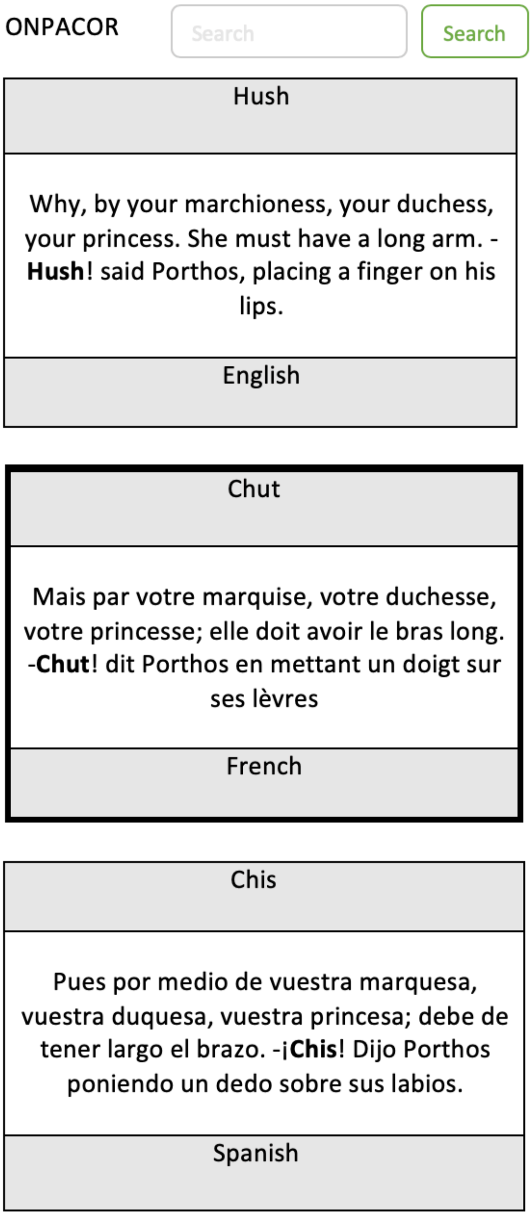


Figure 1: Structure of the database of ONPACOR

The two tables in Figure 1 establish a relationship between an onomatopoeia and one or more concordances, since the onomatopoeic word sometimes coincides in the three languages. Likewise, the concordances are able to establish relationships between themselves, so that the three translations can be related in order to align the extracts in the parallel corpus.

The third step in the implementation of the interface was to set up the website. The site hosting ONPACOR has been created by using *Django*, a Python framework for web development. This website displays two main sections: the administration area and the users' area. The former is used to input the information that is going to be displayed and is accessible only to the creators. A number of safeguards have been put into place when creating a new entry. For example, each language can be added only once, so that no mistake can be made by the administrator. In addition, if the original language option has already been selected, the system will not allow the administrator to add more original languages to that concordance. In the user area, there is a query box to type in the onomatopoeias to be searched for. When typing an onomatopoeic word, the system displays the matches already available in the database to help the user. Afterwards, by clicking on the search button, queries yield concordances in the three languages with a 100-word context, which helps see their meaning and use. The box displaying the excerpt in the original version is highlighted in bold typeface and a thicker frame line. This feature is particularly useful for translators, since it helps them identify clearly which is the original excerpt and which is the translation. In Figure 2, the original text is the French one, extracted from *Les Trois Mousquetaires* by Alexandre Dumas.



Hush

Why, by your marchioness, your duchess, your princess. She must have a long arm. - **Hush**! said Porthos, placing a finger on his lips.

English

Chut

Mais par votre marquise, votre duchesse, votre princesse; elle doit avoir le bras long. -**Chut**! dit Porthos en mettant un doigt sur ses lèvres

French

Chis

Pues por medio de vuestra marquesa, vuestra duquesa, vuestra princesa; debe de tener largo el brazo. -¡**Chis**! Dijo Porthos poniendo un dedo sobre sus labios.

Spanish

Figure 2: Screenshot of the web app ONPACOR

4. Applications of ONPACOR

In our view, ONPACOR has three main applications. Firstly, when finished, it will constitute a huge compilation of literary texts that not only translators but also lexicographers and writers in general can use. Secondly, ONPACOR will become a multilingual resource available for the research community. Thirdly, it will allow researchers to conduct comparative studies of onomatopoeias in different languages and account for the different mechanisms employed by translators.

One of the main reasons for the compilation of ONPACOR was to compare onomatopoeias crosslinguistically. There is evidence that different onomatopoeic words represent the same sound, as the example of the rooster's crow mentioned in Section 2 clearly showed. However, this crosslinguistic variation does not always take place. Generally speaking, the most common strategy used in the translation of onomatopoeias is that of 'equivalence' (Mayoral Asensio 1992: 139). In this sense, the translator uses an onomatopoeic word available in the target language, which is equivalent in meaning to the original one. This is illustrated in the examples provided in Figure 2 (see Section 3.4) where the translators use *hush!* for English, *¡chis!* for Spanish and *chut!* for French.

Besides, another common strategy is the use of loanwords, that is, to use the onomatopoeia in its original form without translating it. Even though the use of loanwords is very common in comics, according to Mayoral Asensio (1992: 139), translators should avoid this strategy which impoverishes the target language and causes onomatopoeic words to disappear due to lack of use.

It is undeniable that, due to their use in comics and because English is the most widely used language on the Internet, onomatopoeias are more frequently attested in English than in Spanish or French. What is more, some of the English onomatopoeic forms have been assimilated in languages such as Spanish and French, as is the case of *boum* or *pum* (de Buron-Brun 2006: 770). This way of proceeding has been found in the translations of *Le Rouge et le Noir* by Stendhal —with the onomatopoeia *bah!*— and in those of the comics of *Astérix et Obélix* —where only the laughter bubbles of *ha ha* were translated into Spanish as *ja, ja, ja*. As the onomatopoeia of laughter coincides in both English and French, its translation was not necessary. Alternatively, as some onomatopoeic expressions are difficult to translate because of lack of correspondence, translators tend to omit them in the target text. Sometimes the omission is caused by the presence of other elements, such as images, which make the onomatopoeias redundant. In our corpus, we found this mechanism, among others, in *The Colour of Magic* by Terry Pratcher, where the onomatopoeic word is found in the English version, but not in the Spanish and the French versions.

5. FURTHER RESEARCH

Once this large trilingual database has been compiled, the next step will be the compilation of a multilingual dictionary of onomatopoeias. The idea is to make use of all the collected data to create an electronic tool that will display the examples and the meaning of each onomatopoeic form. In fact, a tentative version of this dictionary has already been implemented, though only for Spanish (Orrequia-Barea and Marín-Honor 2018). In this dictionary, the user can search either for the onomatopoeic word, for example, *pum*, or for the real sound or noise it represents, which is a gunshot. These two options are necessary, since onomatopoeic expressions are usually associated with more than one sound or vice versa. Apart from the onomatopoeia and its meaning, the dictionary also displays the concordances so that users may see the onomatopoeic forms in context. We are currently working on the English and French versions, but we do not discard the possibility of widening this project to other languages, such as Italian and German.

As regards the translations, we have chosen just one translation for each original book. However, it would be interesting to look into different translations of the same work by different translators. This may help improve our understanding of the mechanisms that are used by translators as well as include variations or different renderings of the same onomatopoeic word.

In addition, in the future the corpus and derived dictionaries may well be complemented by the creation of an extension for messaging apps allowing users to include onomatopoeias in their conversations. Given the widespread use of text messaging today, people need to convey as much accurate information as possible in their conversations to put forward messages properly. It is true that, to a certain extent, this need is already catered for by emojis, stickers and gifs but it is undeniable that these resources do not always serve our communicative needs. That is why we are considering the implementation of an onomatopoeia extension that would allow users to introduce them in the messages without having to actually type them. The idea is to use the previous compilation of onomatopoeic words to create what may be termed ‘onomojis’, as we have coined them, that is, emojis consisting of an onomatopoeia. Their design could be similar to that used in comics, with the typical font, shape and colours of onomatopoeias in this literary genre.

6. Conclusions

The analysis of onomatopoeias has been neglected in linguistics. Adding to the scarce literature on the topic, it has been shown that there is a lack of multilingual resources related to these types of words. For this reason, ONPACOR proves to be a useful digital tool, which can shed light on many onomatopoeias related issues. ONPACOR constitutes a really useful resource for translators, filling the gap that currently exists in the multilingual dimension of onomatopoeias.

For the sake of accuracy, and to include just onomatopoeias in ONPACOR, a thorough study has been carried out regarding the differences and similarities between onomatopoeias and interjections. This means that translators are going to have a solid compilation of onomatopoeias which is going to help them in their translations. The interface displays the onomatopoeic words in the three languages, and provides information about the context which, in turn, may also help the translator decide whether the onomatopoeia fits in the translation. Additionally, the concordances for the immediate context and the translations have been carefully selected so that they represent evidence of real language in use. Furthermore, the interface will allow translators and linguists to conduct comparative studies between different languages and to carry out research into the strategies deployed by translators. Although the web app has already been created, this is still an on-going project and we are currently introducing new onomatopoeias and concordances as the project progresses.

References

Anthony, Laurence. 2019. *AntConc*. Tokyo: Waseda University.

ATILF. *Base Textuelle Frantext.* 2019. ATILF-CNRS and Université de Lorraine. https://www.frantext.fr/

Arboleda, Ana M. and Carlos Arce-Lopera. 2017. The French, German and Spanish sound of eating fresh fruits and vegetables. *Food Research International* 102: 171–175.

Assaneo, María Florencia, Juan Ignacio Nichols and Marcos Alberto Trevisan. 2011. The anatomy of onomatopoeia. *PLoS ONE* 6/12. https://doi.org/10.1371/journal.pone.0028317 (20 April, 2020.)

Barbéris, Jeanne-Marie. 1992. Onomatopée, interjection: Un défi pour la grammaire. *L'Information Grammaticale* 53/1: 52–57.

Boole, George. 1854. *An Investigation of the Laws of Thought. On which are Founded the Mathematical Theories of Logic and Probabilities*. New York: Dover Publications.

Bueno Pérez, María Lourdes. 1994. La onomatopeya y su proceso de lexicalización: Notas para un estudio. *Anuario de Estudios Filológicos* 17: 15–26.

Chen, Peter Pin Shan. 1976. The entity-relationship model: Towards a unified view of data. *ACM Transactions on Database Systems (TODS)* 1/1: 9–36.

Coronel, Carlos and Steven Morris. 2016. *Database Systems: Design, Implementation and Management*. Boston: Cenage Learning.

Corpus de Referencia del Español Actual. 2020. Real Academia Española: Banco de datos (CREA). http://www.rae.es.

de Buron-Brun, Bénédicte. 2006. La onomatopeya, ¿mucho ruido para pocas nueces o un rompecabezas para el traductor? In Manuel Bruña Cuevas, María de Gracia Caballos Bejano, Inmaculada Illanes Ortega, Carmen Ramírez Gómez and Anna Raventós Barangé eds. *La Cultura del Otro: Español en Francia, Francés en España*. Sevilla: Universidad de Sevilla, 768–784.

de la Rosa Regot, Nuria. 2015. *Translating Sounds: The Translation of Onomatopoeia between English and Spanish*. Barcelona: The University of Barcelona Degree Dissertation.

de Saussure, Ferdinand. 2011. *Course in General Linguistics*. New York: Columbia University Press.

Diccionario de la Lengua Española. 2020. Real Academia Española. https://dle.rae.es.

Enckell, Pierre and Pierre Rézeau. 2003. *Dictionnaire des Onomatopées*. Paris: Puf.

Gasca, Luis and Román Gubern. 2008. *Diccionario de Onomatopeyas del Cómic*. Madrid: Ediciones Cátedra.

Husillos Ruiz, Araceli. 2018. *Las Pequeñas Palabras y su Traducción: Glosario Trilingüe de Onomatopeyas*. Valladolid: The University of Valladolid Degree Dissertation.

Kleiber, Georges. 2006. Sémiotique de l’interjection. *Langages* 40/161: 10–23.

Kwon, Nahyun. 2015. *The Natural Motivation of Sound Symbolism.* Australia: Queensland University PhD Dissertation.

Mayoral Asensio, Roberto. 1992. Formas inarticuladas y formas onomatopéyicas en inglés y español. Problemas de traducción. *Sendebar: Revista de La Facultad de Traducción e Interpretación* 3: 107–140.

Melnikienė, Danguolė. 2016. Le statut grammatical des onomatopées dans la linguistique moderne. *Verbum* 6/6: 168–187.

Merriam Webster Dictionary Online. 2020. https://www.merriam-webster.com

Orrequia-Barea, Aroa and Cristian Marín-Honor. 2018. Hacia la elaboración de un diccionario de onomatopeyas en español. *CHIMERA: Romance Corpora and Linguistic Studies* 5/1: 93–99.

Rhodes, Richard. 2010. Aural images. In Leanne Hinton, Johanna Nichols and John J. Ohala eds. *Sound Symbolism*. Cambridge: Cambridge University Press, 276–292.

Sugahara, Takashi. 2011. *Onomatopoeia in Spoken and Written English: Corpus and Usage-based Analysis*. Japan: Hokkaido University PhD Dissertation.

Swiatkowska, Marcela. 2000. *Entre Dire et Faire. De l’Interjection*. Krakow: Wydawnictwo Uniwersytetu Jagiellonskiego.

The British National Corpus 3.0. 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

Corresponding author
Aroa Orrequia-Barea
Univesity of Jaén
Department of English Studies
Building D2, office 241
Campus Las Lagunillas
23071, Jaén
Spain
E-mail: orrequia@ujaen.es

received: February 2020
accepted: June 2020

RiCL Research in Corpus Linguistics

Grammaticalisation paths in the rise and development of *aside*

Rodrigo Pérez Lorido – Pablo Ordóñez García
University of Oviedo / Spain

Abstract – In this paper we analyse the grammaticalisation processes involved in the rise and development of the *a*-adverbial *aside* from the original combination of the preposition *on* and the substantive *side* in Old English. Different aspects of this grammatical change will be discussed in the paper, from morphosyntactic and phonological (coalescence-univerbation) to semantic ones (development of abstract senses, extension of semantic range), taking very much into account the diachronic axis that underpins them. Special attention has been paid in the analysis to the variation patterns of *aside* that existed in the Late Middle English period (when the actual process of grammaticalisation was about to be completed) and to the correlation of these variants with the geographic provenance of the texts, trying to determine if the processes of word formation that gave rise to this new word class travelled homogeneously across Britain.

Keywords – *aside*; *a*-adverbials; grammaticalisation; decategorisation; coalescence; attrition

1. Introduction: Grammatical status and origin of *aside*

The status of the English word class of elements beginning with *a-* like *aside*, *ahead* or *anew* is a complex question that has constituted "a problem in classification for grammarians" (Quirk *et al*. 1985: 408) for a long time. Thus, some *a*-prefixed words in English have been classified as adverbs by Quirk *et al*. (1985: 408–409, 516) but as prepositions by Huddleston and Pullum (2002: 613–614), to which we may add the fact that other *a*-words in English like *asleep* or *alive* also display the grammatical characteristics of adjectives. Regarding *aside* specifically, whereas Schlütter (2008: 149) and the *Oxford English Dictionary* (OED) take its adverbial status for granted, the word is classified as a preposition by Huddleston and Pullum (2002: 614). One central aspect of Huddleston and Pullum's argumentation is that *aside*, with the meaning 'not including, except',[1] can take a complement (a *from*-phrase), in which case the preposition precedes

[1] Let's remember that this usage is characteristic of standard American English. The British English choice of preposition for this sense is *apart*.

Research in Corpus Linguistics 8/2: 63–86 (2020). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.08.02.04

its complement (1a), or even an NP complement, where the preposition exceptionally follows it (1b).

(1a) Aside from carrots, my daughter won’t eat vegetables.

(1b) Carrots aside, my daughter won’t eat vegetables.

Additionally, the fact that *aside* (with the meaning ‘apart, to the side’) is mostly used in English as an adjunct that modifies a verb without an NP complement is not —according to Huddleston and Pullum (2002: 612–614)— determinant to label it as an adverb, since other prepositions perform this function in a similar way (see (2)):

(2) They put the documents aside. / They put the guests up.

Huddleston and Pullum (2002: 613) also point out to the exceptionality of *aside* within the group of *a*-prepositions as —unlike most of them— it may occur as goal complement with verbs of motion (3) but not as locative complement of *be* (4):

(3) They went abroad/ashore/adrift. / They stepped aside.

(4) They are abroad/ashore/adrift. / *They are aside.

Another exceptional feature of *aside*, according to Huddleston and Pullum (2002: 614), is that, in its spatial sense, it usually occurs only in dynamic contexts (5). Note, however, the stative context in (6):

(5) He pushed them aside. / *They are aside.

(6) He pushed them aside. / They stood aside.

Notwithstanding Huddleston and Pullum’s (2002: 212–214) argumentation, a majority of grammars, diachronic corpora and historical dictionaries of the English language place *aside* within the set of English adverbs (see, for instance, the OED, the *Middle English Dictionary* [MED] and Quirk *et al.* 1985: 1151). Consequently, in this paper we will refer to *aside* as an *a*-adverbial and as a representative of that word class, irrespective of the fact that it clearly developed prepositional functions along its history. Moreover, it must also be remembered that other *a*-words like *asleep*, *alive* or *afraid* are traditionally categorised as adjectives and not as adverbs. Among the differences between the former and the latter, Quirk *et al*. (1985: 408–409) point out that *a*-adjectives may function both predicatively and attributively, though the latter only marginally, and when modified, (see (7)). In turn, *a*-adverbs can only be used predicatively (see (8)).

(7a) The children were *asleep.*

(7b) The *fast asleep* children.

(8a) The ship was *ahead.*

(8b) *An *ahead* ship.

In this respect, *aside* clearly falls within the range of *a*-adverbs, as it can never be used attributively.

Concerning the question of the origin of *a*-adverbials/adjectives, the standard assumption is that they are mostly the result of a process that combines, on the one hand, coalescence (fusion) of a preposition with a following complement and, on the other, attrition (phonological erosion) whenever the preposition itself is not *a* (e.g. *ashore* < eMoE *a shore*; *afoot* < ME *a fote*; *aback* < OE *on bæc*; *asleep* < OE *on slæpe*; *anew* < OE *of niowe*).[2] In some cases, when the first element is a prefix (such as OE *ge-*), the process involves attrition alone (e.g. *aware* < OE *gewær*; *ashamed* < *gesceamod*). However, with the exception of the OED, the picture provided by reference grammars and dictionaries about the diachronic development of *a*-words in English is in general fragmentary and imprecise, for three main reasons: 1) a consistent periodisation of the changes is missing; 2) examples of *a*-adverbials involving coalescence of *a* with a following stem are often lumped together with examples which involve more complex derivational processes of coalescence and attrition with other preposition and prefixes; 3) in general the role of the analogical pressure exerted by some forms upon others is underplayed or straightforwardly dismissed.

Regarding *aside* in particular, Huddleston and Pullum (2002: 614) remark that

> [*aside*] contains the prefix *a-*, which originates historically in a form of the preposition *on* [and] is the result of fusion of the preposition with its complement.

[2] Kemenade and Los (2003: 86) also remark that “many OE particles are the result of the grammaticalisation of a PP (see, e.g., *adun* ‘down’ < OE *of dune* ‘off the hill or height’ (OED s.v. *down*, adv.), *aweg* ‘away’ < *on weg* ‘on one’s way’ (OED s.v. *away*, adv.)).”

This view is shared both by the OED, where the etymology of the entry *aside* reads ‘originally a phrase: *on side*’ (OED s.v. *aside* adv., prep., adj. and n.), and the MED, which likewise presents the etymology of *aside* as coming from *on side* (MED s.v. *asīde* adv.). In other sources, such as *The Random House Dictionary of the English Language* (Random House 1987: 1), reproduced verbatim in several online dictionaries such as Dictionary.com[3] or Collins,[4] the emphasis is placed on *aside* as containing the preposition *a*, which ultimately derives from a reduced form of *on*:

> [*a*-] a reduced form of the Old English preposition *on*, meaning “on”, “in”, “into”, “to”, “toward”, preserved before a noun in a prepositional phrase, forming a predicate adjective or an adverbial element (*afoot*; *abed*; *ashore*; ***aside***; *away*) [our emphasis].

This is an example of the inaccuracy of some accounts of *a*-words in English, as this source puts together elements that truly descend from ‘*on* + NP’ combinations (*afoot*, *abed*, *aside* and *away*) with others whose origin is different: there is no clear evidence for the origin of *ashore* as ‘on shore’. The OED suggests, rather, ‘*a* + *shore*’ as the correct source (OED s.v. *ashore*, adv).

2. Aims and scope

As stated before, the study of *a*-prefixed elements in English is characterised by a certain conceptual and terminological fuzziness, as well as by a clear focus on the synchronic side of the problem. Much of the discussion about *a*-words in English has centred on their typological status, leaving aside the analysis of the diachronic processes that gave way to this word class and the multiple interactions among them. We can say, therefore, that there is no single comprehensive, wide-ranged corpus study of *a*-adverbials/*a*-adjectives in English to the present date aiming at providing a clear perspective on their rise and development. In order to bridge this gap, this paper aims firstly at providing *prima facie* evidence about the abstract grammatical mechanisms which prompted the rise and development of the *a*-adverbial *aside* in English (especially insofar as grammaticalisation is involved), and, secondly, at offering a detailed historical overview of the rise and consolidation of this *a*-word in the history of English. A special point is made in the study of analysing the patterns of variation for the different forms of *aside* in the late Middle

[3] https://www.dictionary.com/browse/a-

[4] https://www.collinsdictionary.com/dictionary/english/adoze

Ages according to their geographical distribution, inasmuch as they may provide evidence about how this development travelled across Britain. For that purpose, an extensive corpus of Old and Middle English texts has been surveyed, trying to make up for the lack of quantitative support which underlies most analyses of *a*-elements in English today.

3. DATABASE

The data for the study was drawn from two computerised corpora: the *Dictionary of Old English Corpus* (DOEC), which was examined in its entirety for the OE section of this work, and the *Corpus of Middle English Prose and Verse* (CMEPV), from which 63 entire texts[5] were analysed for the ME section. The DOEC is the most comprehensive digitised corpus of OE prose and verse to date, containing 3,037 texts amounting to more than 3 million words. It contains a copy of every extant text in OE, including minor works such as wills, riddles, glosses and charters. The CMEPV is the largest collection of complete ME texts available online, containing 146 items, mostly focused on the Late Middle English period (LME). The data for the OE section of the study was retrieved manually and the data for the ME section was retrieved using the concordancer provided by the CMEPV. Apart from the DOEC and the CMEPV, the *Linguistic Atlas of Late Medieval English* (eLALME) and historical dictionaries such as the OED and the MED have also been employed as data sources.

4. ANALYSIS

4.1. The change on side > aside*: Grammaticalisation and semantic change*

As mentioned in Section 1, the standard assumption about the origin of *aside* is that it arose from the combination *on side* in OE. However, according to the OED and the MED, the first attested instances of *aside* expressing an abstract notion of distancing or detachment ('off to one side, out of the way') in the sense of the modern adverb date back to 1330, as shown in (9) and (10) below.

[5] The complete list of the texts analysed can be found in the appendix to this study.

(9) Þe sarrazins seiȝe þai com & flowen **oside** [Cai: asyde], alle & som.[6]
'The Saracens see them come and flee aside, all together'
(MED c1330 (?a1300) *Rich.* (Auch) 119/130)

(10) Otuwel starte **o side**, & lette þe swerd bi him glide.
'Otuel jumped aside and let the sword glide about him'
(MED c1330 *Otuel* (Auch) 537)

Other early instances of *aside* referred to in the OED and the MED, always involving dynamic contexts, are illustrated in (11) and (12) below.

(11) Þe coupes of gold were treden **a-syde** al with mannis fet.
'The golden cups were pushed aside all with men's feet'
(OED, c 1380 Sir Ferumb. 2297)

(12) His hors for feere gan to turne and leep **asyde** and foundred as he leep.
'His horse, for fear, began to turn and leapt aside and stumbled as he leapt'
(MED c1385 Chaucer CT.Kn (Manly-Rickert) A.2687)

Our analysis of the OE corpus, however, suggests that the traditional chronology for the rise of *aside* as an adverbial element meaning 'off to one side' is perhaps erroneous, and that the beginning of the process of grammaticalisation and semantic change that led from the combination *on side* meaning 'on the side of the body' to 'off to one side, out of the way' (either in a concrete or an abstract sense) should be pushed back in the history of English at least 300 years, within the OE period. Let us see to this development in detail.

As pointed out by Rissanen (2004: 154), *side* was used in OE in its basic notion 'side of human or animal body'[7] in the vast majority of cases. Our own analysis reveals that out of more than 283 examples of *side* in the DOEC, 208 have this meaning, confirming Rissanen's view (see (13) and (14)):

(13) **His side** wæs on ðære rode gewundod.
his side was on the rood wounded
'His side was wounded on the rood'
(DOEC ÆCHom II, 12.1)

[6] Gonville and Caius College MS 175/96, on which Karl Brunner (1913) based his edition of the ME romance of Richard Lionheart, has *þe Sarezynes seyȝen þat þey come, and ffleyȝ asyde, alle and somme*.
[7] *Side* is also used metaphorically in OE to refer to parts of ships, mountains or buildings.

(14) **Wiþ** **sidan** **sare** genim þære ylcan wyrte...
against of-the-side pain take the same herb
‘Against pain in the side, take the same herb…’
(DOEC Lch I (Herb))

The next stage in the semantic development of *side* in the OE period involved signalling position (usually ‘immediate neighbourhood’) with regard to another element, or direction (‘to this or that side’),[8] which marks the first step toward the generalisation of the meaning of the word and the future direction of grammaticalisation paths. In these contexts, *side* is usually preceded by a preposition, normally *at* or *on*, as shown in (15) and (16) respectively:

(15) þa ne wiste he hwæt he gefelde cealdes **æt his sidan** licgan.
then not knew he what he felt of-cold at his side lie
‘Then he didn’t know what cold thing he felt lying by his side’
(DOEC Bede)

(16) þeos ðridde india hæfð **on anre sidan** þeostru & on oþre ðone grimlican
this third India has on one side darkness and on other the fierce
garsecg.
ocean
‘This third of India has darkness on one side and the fierce ocean on the other’
(DOEC ÆCHom I, 31)

Then, the cognitive polysemy of *on* allowed for a commonly observed move in the development of spatial meanings for prepositions, namely, changing their domain of reference from the idea of ‘contact’ to ‘extending beyond’ along certain orientational axes, as in the case of *over* (Tyler and Evans 2003: 78–84; Brenda 2014). As regards the combination *on side*, Figure 1 below tries to represent the process which involved the change in the referential domain of *on* from ‘in the immediate neighbourhood of’ or ‘adjacent to’ (A) to ‘off to one side’, with an implication of distancing (B):

[8] Bodily parts are, as an anonymous reviewer of this text remarked, commonly used across languages to conceptualise spatial relationships (see Heine *et al.* 1991: 34 and Heine and Kuteva 2002).

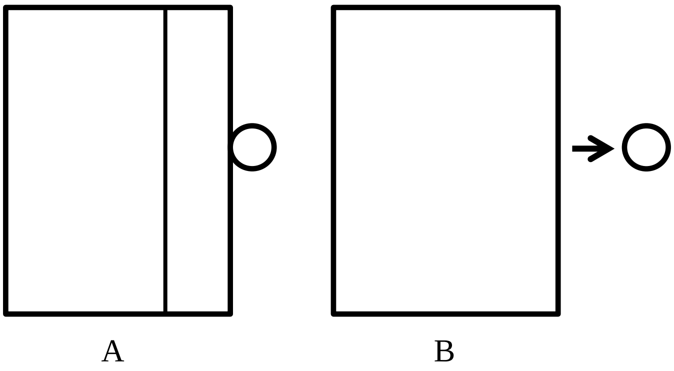


Figure 1: Spatial scene involving the transition from 'adjacent to' to 'away from' for *on* in *on side*

In this process, subjectification (a changing point of view, from an outsider's to the referent itself) may probably have played a role,[9] in a similar way to the one suggested by Rissanen (2004) for the extension of the meaning of *side* from nearness to distancing in the adverbial *besides*:

> As long as the point of view is an outsider's or neutral, the idea of 'side' most naturally refers to somebody or something in the immediate vicinity of the person or object governed by *beside(s)*. But when the relation is defined from the point of view of its referent, distancing, movement away, becomes a natural extension of meaning. With this development the way is paved for the emergence of abstract meanings. (Rissanen 2004: 162)

In any case, the semantic change undergone by *on side*, from signalling position to indicating 'detachment' or 'distancing' (characteristic of an incipient grammaticalisation stage) would not have taken place yet in the OE period, according to the OED and the MED. This assumption is, however, not borne out by our corpus analysis, as mentioned at the beginning of this section: after checking all the instances of *side* preceded by *on* in the entire DOEC, it appears that at least one instance of *on side* had the adverbial meaning 'off to one side, aside', with a clearly implicit sense of detachment. It is Riddle 21, from the *Old English Riddles*, whose solution is 'the plough' (see (17)).

(17) Fealleþ **on sidan** þæt ic toþum tere,
falls on side that I with-teeth tear
'What I tear with my teeth falls aside/to the side'
(DOEC Rid 21)

It is true that this is the only example of the combination *on side* in the entire DOEC with that sense, but it is revealing enough to suggest that grammaticalisation was perhaps on

[9] See Traugott (1999) for an interesting discussion of the overall influence of subjectification in grammaticalisation phenomena.

its way already in OE and to consider rethinking the chronology of the entry *aside* in the historical dictionaries of the English language.[10] On the other hand, it is worth noting the relevance of the collocation *on sidan*, in which the preposition and the substantive are adjacent, which provides a natural transition to the univerbated forms *oside*, *osyde* characteristic of the more grammaticalised forms of the Middle Ages. That said, the combination *on side*, with strict adjacency of both words, was very infrequent in OE (only three examples out of 238 instances containing the word *side* in the entire DOEC). Apart from our example (17), the other two are shown in (18a) and (18b) below, none of which has the meaning 'off to one side, out of the way'.

(18a) ic geseo þa dolhswaðu on his handum. & on fotum. & **on sidan**.
I see the scars on his hands and on feet and on side
'I see the scars in his hand and in his feet and in his side'
(DOEC ÆCHom I, 16)

(18b) Oft **mec** isern scod sare **on sidan**.[11]
often me iron hurt painfully on side
'Often iron hit me painfully on the side / hit my side painfully'
(DOEC RId 72)

Actually, the majority of examples of *side* preceded by *on* in the corpus (37 out of 238 instances) have another intervening element between the preposition and the substantive (typically the feminine determiner that goes along with *side*) and basically refer to the side of the human body as a reference point, rather than to movement away from it (see (19a) and (19b).

(19a) Ða ða heo geseh niman hire cild & […] mid spere gewundian **on ða sidan**…
when she saw take her child and with spear wound on the side
'When she saw (them) take her child and wound (him) in the side…'
(DOEC ÆHom 12)

[10] This is in contradiction with the opinion of one anonymous reviewer of this paper, for whom OE shows no trace of the grammaticalisation of *side* at all. We think, however, that this example is convincing enough.

[11] This example is reminiscent of external possession structures in OE (commonly used when referring to parts of the body and to inalienable possession in general), in which the possessive relationship is expressed, not by a phrase containing a possessive adjective *on side min*, but by an NP usually in the dative case (although *mec* is in the accusative case here) acting independently as experiencer of the action plus an NP or PP standing for the possessed item (in this case *side*). See Pérez Lorido and Casado Núñez (2017) and, more recently, Allen (2019) for a detailed account of external possession in OE.

(19b) wið sidan sare, rudan wið rysele gemenged […] lecge **on þa sidan**.
against of-the-side pain rue with lard mixed lay on the side
‘Against pain in the side, lay rue mixed with lard on the side’
(DOEC Lch II (1)) [027600 (21.1.5)]

A summary of the general incidence of *side*, *on ... side* and *on side* in the DOEC can be seen in Table 1 below.

side, *on… side* and *on side* in the corpus	Raw figures
side	**283**
on ... side	**37**
on side	**3**

Table 1: Incidence of *side*, *on ... side* and *on side* in the DOEC

It must be noted that in none of the Old or Middle English examples presented so far was *on side/aside* used in clearly abstract contexts. The first examples of a truly metaphorical or figurative use of *aside* as ‘putting or setting aside feelings, attitudes or emotions’ are found at the end of the LME period, in any case not earlier than 1390, according to the OED and the MED, as shown in (20).

(20a) Wyues..moste..at nyght..**leye** a lite hir holynesse **asyde**.
‘Women, mostly at night, lay their holiness aside a little’
(MED c1390 Chaucer *CT.ML*) Manly-Rickert, B.713)

(20b) For al this pompe and al this pride, Let no justice **gon aside**.
‘Let no justice go aside, for all this pomp and pride’
(MED ca1393 Gower CA (Frf3) 7.2388)

(20c) Al fer and drede was **leide a-syde** & goon.
‘All fear and dread were laid aside and gone’
(MED c1425 Lydg. *TB* (Aug A.4), 1.3337)

This implies a further step in the grammaticalisation of *on side*, representing an even greater generalisation of the meaning of *side* (from concrete to abstract), which will run in parallel with other morphosyntactic and phonological processes of coalescence and attrition to various degrees throughout the period, as will be shown presently (see Section 4.2). The last stage in the grammaticalisation of *on side*, however, was the incorporation of *aside* to the category of prepositions in the Early Modern and Modern English periods, which completed the cycle from a fully denotative, semantically rich lexical item into a grammatical marker (Hopper and Traugott's 2003: 106–115 'decategorisation'). Interestingly, the use of *aside* as a preposition did not initially take a *from*-phrase complement with the meaning 'except' or 'apart from', as in (1a) above, but an *of*-phrase with the meaning 'alongside, by the side of' (21), or an NP complement with either the meaning 'at the side, beside' or 'past, beyond', as in (22) and (23) below:

(21) A shippe […] which tooke his course **aside of vs**.
(OED, 1630 Wadsworth Sp. Pilgr. iv. 33)

(22a) And in the ashes sat, **aside the fire**.
(OED, 1615 Chapman Odyss. Vii. 215)

(22b) The shop that was **aside the house**.
(OED, 1743Wesley Wks. (1872) XIII. 175.)

(23a) The kind Prince, Taking thy part, hath rusht **aside the Law**.
(OED, 1592 Shakes. Rom. & Jul. iii. Iii.26)

(23b) Which resolution he had taken up before … and was put **aside it**, by the amplitude of that Fortune.
(OED, 1663 Flagellum or O. Cromwell (1672) 22)

The prepositional use of *aside* in the collocation *aside from*, meaning 'except' (first recorded in 1818 and used primarily in standard American English) represents the final stage in the semantic development associated with the grammaticalisation of the combination *on side*, fully taking on an abstract sense of *side* and encroaching metaphorically[12] on the idea of detachment (see (24a) and (24b) below):

(24a) **Aside from this**, the mere show is more magnificent than can be seen at any other court in Europe.
(OED, 1818 Ticknor in Life, Lett., & Jrnls. 206)

(24b) The college […] possesses revenues, **aside from tuition**, sufficient to maintain the faculty.
(OED, 1847 L. Collins Kentucky 507)

[12] See Lakoff and Johnson's (1980: 14-21) notion of orientational metaphors in this respect.

Summarising, the history of the semantic development of *on side* associated to its grammaticalisation involves the shift of the meaning of *side* from a concrete, stative meaning (‘part of the human or animal body’) to senses indicating position (‘at the side’), to later develop a sense of distancing/detachment (‘off to one side’) either physically or metaphorically, to finally signal purely abstract reference (‘aside from…’). Linguistically, we also observe a characteristic cline in the grammaticalisation of *on side* > *aside*, this is, the shift of an independent word belonging to a major lexical category (the substantive *side*) to the status of grammatical element marking a particular construction (the preposition *aside*). In the following section we will analyse the morphosyntactic and phonological processes which run in parallel with the semantic changes discussed so far.

4.2. The change on side > aside*: Coalescence and attrition*

Most accounts of grammaticalisation (Lehmann 1985; Croft 2000; Heine 2003; Hopper and Traugott 2003) assume that items subjected to that process become in general less autonomous. Autonomy is measured, according to Lehmann (1985: 305), by three parameters: weight, cohesion and variability. Each of these parameters have a paradigmatic and a syntagmatic dimension, which relate to the selectional aspects of the sign and to its combinatorial potential respectively. A gain in bondedness (the syntagmatic aspect of cohesion) is usually referred to as ‘coalescence’ (Lehmann 1985: 308), and manifests itself as syntactic elements becoming morphological, while a loss in paradigmatic weight is conventionally named ‘attrition’ (Lehmann 1985: 307), and results in the gradual loss of phonological or semantic substance. In this section, the morphosyntactic and phonological processes of coalescence and attrition that affected the change *on side* > *aside* are analysed, with a focus on the LME period. The reason why emphasis in this part of the study is laid on the LME period is that this stage of the history of English is more likely to reflect the consolidation of the processes of coalescence and attrition involved in the grammaticalisation of *on side* while also displaying a remarkable degree of dialectal variation than the Early Middle English period (EME). This makes

LME a better candidate for producing relevant evidence about both the diachronic evolution of *on side* > *aside* and its geographical distribution.[13]

The methodology employed in this part of the research consisted of analysing 63 LME texts copied in the late fourteenth and fifteenth centuries as contained in the *Corpus of Middle English Prose and Verse* (CMEPV), recording the different spelling variants of *on side*/*aside* and mapping them against different variables. These included the degree of bondedness and attrition between both parts of the original structure, the date of the manuscripts[14] and the geographical provenance of the texts. The texts were scanned and the data retrieved using the search engine and concordance software provided by the corpus itself.

The complex processes of linguistic change and word formation which led from OE *on side* to Present-Day English (PDE) *aside* mentioned in the previous section caused a considerable number of spelling variants to appear in ME. The list of such variants recorded in our database amounts to nine. These are: *aside*, *asyde*, *on syde*, *on side*, *asides*, *on syd*, *asydis*, *azide* and *oside*. All of them are present in the OED entry for *aside,* which additionally includes the variants *acyde* and *assyde*, and in the MED. As we can observe, the spelling variants in our corpus fall under two major categories according to the degree of bondedness between both parts of the original structure *on side*, from the more to the less cohesive: 1) amalgamated forms (*aside*, *asyde*, *asides*, *asydis*, *azide* and *oside*) and 2) separated ones (*on syde*, *on side* and *on syd*). It must be remembered, however, that some of the spellings for *aside*/*asyde* in the electronic editions from which the ME data for this work have been retrieved present hyphenated forms (*a-side*/*a-syde*) or a physical

[13] It must be agreed, however, as two anonymous reviewers of this work pointed out, that an analysis of the EME texts would be desirable in order to obtain a clearer picture of the development *on side* > *aside* in the history of English.

[14] The conclusions in this respect will necessarily be limited, as the temporal range of the texts collected for analysis from the CMEPV (1375 to 1475) does not allow for a detailed, fine-grid diachronic analysis. On the other hand, finding out about the exact date of the texts was in some cases very difficult to assess, if not impossible. As a norm we used the characteristics of the manuscripts provided by the references in each of the texts of the CMEPV. In many cases, however, additional research on the date (and place) of the manuscripts was necessary, as the information in the CMEPV was sometimes incomplete or simply non-existent. This was implemented as follows:

i. Consulting the electronic version of the eLALME, which contains relevant information on the place in which most of the texts in the CMEPV were written and the date, or the approximate date, in which they were written.
ii. Consulting the British Library section on manuscripts, the online catalogues of the libraries in which some of the manuscripts were kept and the introductory sections of the editions of some of the texts detailed in the CMEPV. This second option was chosen whenever the information in the eLALME was non-existent or insufficient.

separation between the preposition and the substantive (*a side*/*a syde*). These are, however, editorial conventions adopted in the modern editions from which the electronic texts were transcribed, and do not necessarily reflect the reality of the manuscripts. Particularly, the choice of spacing between *a* and *side/syde* depends very much on editorial practice and is virtually impossible to certify without accessing the manuscripts. Hyphenation, on the other hand, is systematically not found in ME manuscripts, so all examples in the corpus involving it must be considered editorial. Therefore, for the sake of consistency and coherence, all examples of *a-side/a side* and *a-syde*/*a syde* in the corpus have been put together with the instances of *aside*/*asyde* respectively in the quantitative and qualitative analyses.[15]

Regarding frequency, the most frequent type in our corpus is *aside* (217 tokens out of a total of 402), with *asyde* (148 tokens) ranking second, both together representing more than 90% of the total number of tokens in the corpus (90.7%). These two types are immediately followed in frequency by the two-word, more conservative forms *on syde* (22 tokens), *on side* (6 tokens), and *on syd* (2 tokens), which make up 7.4% of the total number of tokens in the corpus. The remaining types can be considered marginal as they represent only 1.7% of the total number of tokens in the corpus, and they all (except the type *oside*) have the vowel *a* as the prefixal element. A summary of the data is presented in Table 2 and Figure 2 below.

Type	**Tokens**	**%**
aside	217	53.98%
asyde	148	36.81%
on syde	22	5.47%
on side	6	1.49%
asides	3	0.75%
on syd	2	0.50%
asydis	2	0.50%
azide	1	0.25%
oside	1	0.25%
TOTAL	402	100%

Table 2: Number of tokens and percentages for the different types of *aside*

[15] We wish to thank an anonymous reviewer of this paper for drawing our attention to this very important fact.

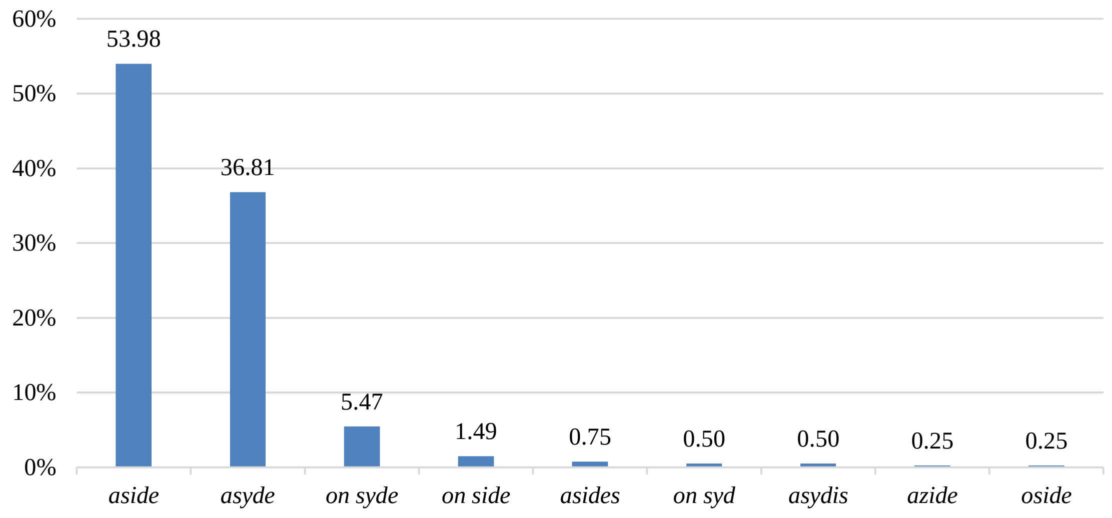


Figure 2: Incidence of the different variants of *aside* in the corpus

We can observe from these figures that the preferred late medieval forms for the original OE combination *on side* were the univerbated ones *aside*/*asyde* (365 out of 402 tokens, which represent 90.7% of the total number of tokens in the corpus). From a diachronic point of view, the corpus analysis confirms that single-word forms systematically outnumber two-word ones at every stage of the development in the period 1375–1475. This clearly reflects the consolidation of the grammaticalisation processes discussed in the previous section in the late Middle Ages. In addition to this, the variation patterns in the corpus data clearly show that the process of attrition of the OE preposition *on* into *a/o* was well under way by the late fourteenth and fifteenth centuries in England, in agreement with Mossé's (1952: 37) statement about vowels in unaccented prefixes like OE *of*, *on*, *on*- and *ond*- having undergone reduction and already turned into *a*- by that period. Indeed, 6 out of the 9 types in the corpus show a reduced form of the original preposition *on* (massively *a*), which turns into 372 instances with reduced vowels in terms of tokens (92.5% of the total). Interestingly, it is only the types in the corpus that display a physical separation between both elements of the original construction that reproduce the preposition *on* in its full form (*on side*, *on syde*, *on syd*).

Turning now to the geographic distribution of the different variants of *on side/aside* in the corpus, one of the goals of this study was —as mentioned before— to be able to associate the different variants to a specific location, so as to establish correlations between area and type which might provide clues about the progress of the change *on side* > *aside* in English. For that purpose, we created a dot map illustrating the location of the different types of *aside* used during the period under analysis, including all the British and Scottish counties, which might provide a visual aid to observe the distribution of the types and the frequency in each area or dialect (Figure 3 below). It is important to remark, however, that only those types of *aside* that were the most used in each manuscript or by each author were included in the map, which explains why, for instance, only seven out of the nine types analysed actually appear in the graph. This means that the ones omitted were either not the predominant types for a given author or in a given manuscript in a specific area, or that their presence was minimal in their respective manuscripts, as in the case of *asides* and *asydis* with 3 and 2 tokens, respectively. Therefore, the dot map in Figure 3 below only displays the most frequently used types by author or manuscript within the different English and Scottish counties.[16]

Otherwise, each of the types present in the map is represented by a symbol with a different hue of black and white. Those in grey colour correspond to the types that reflect the more innovative traits in the spelling (univerbated forms with the former preposition *on* reduced to *a*). The ones in black are those that represent more conservative tendencies, with the preposition *on* separated from the substantive *side*/*syde*. The rest of the types (in white colour) correspond to marginal or minor types. Finally, it is worth mentioning that whenever two symbols are very closely placed on the map, that means that they appeared in the same manuscript or that they were the most predominant types for that specific author.

As can be perceived, grey is virtually the only colour present in the area south of river Thames that corresponds to the Southern and Kentish dialects, and it is also the predominant colour in the Midlands, excluding the Northern Midlands. This fact suggests that the change *on side* > *asid*e perhaps developed faster in those areas (south and south-eastern England as well as in the London area, where the variants *aside*/*asyde* are systematically recorded), or even originated there, although this is difficult to assess without carrying out a thorough analysis of the process in a substantial corpus of EME texts and taking a look at the development of other *a*-adverbials as well.

[16] All the manuscripts that conform the corpus were written in England, except one text of Scottish origin.

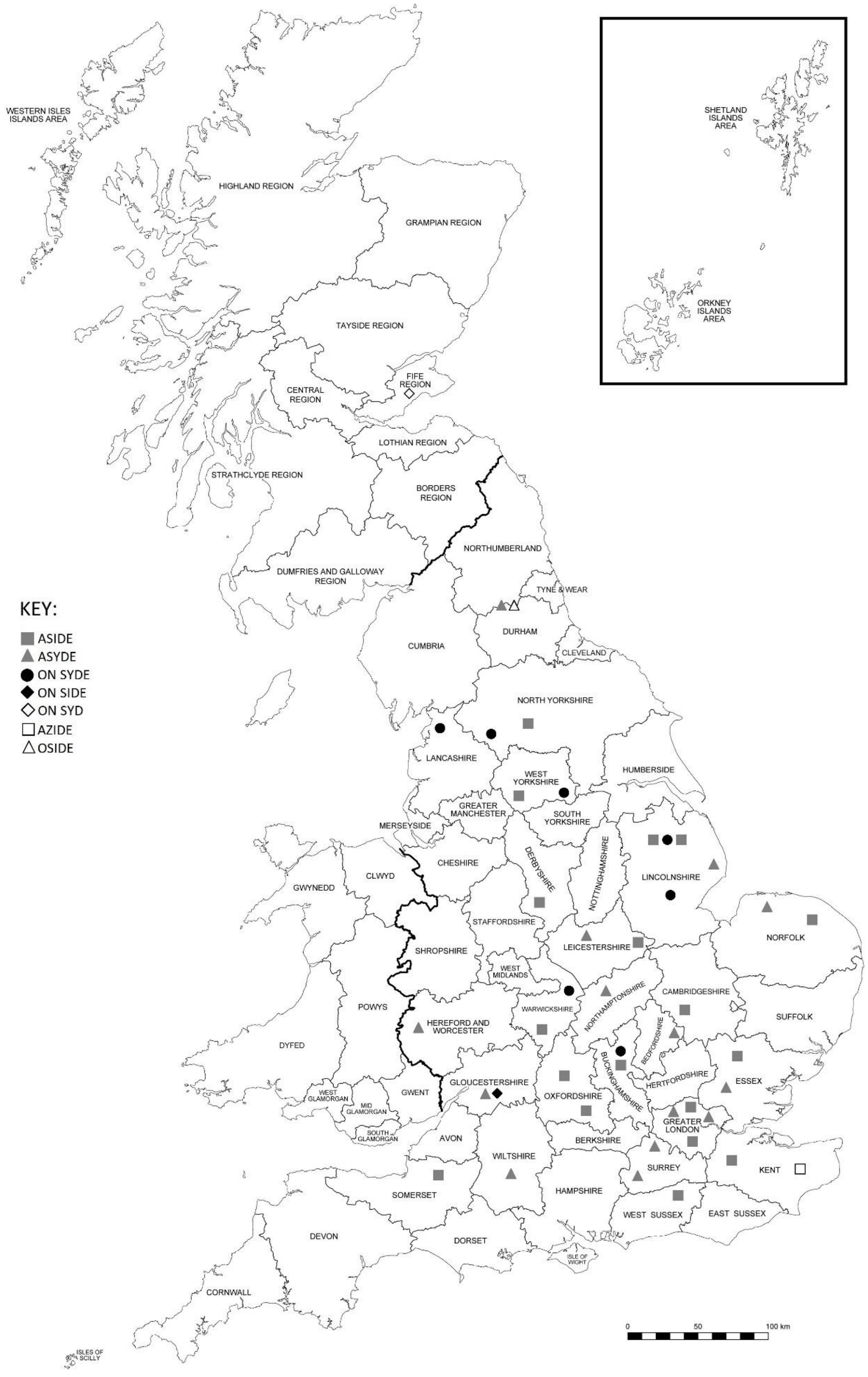


Figure 3: Geographic distribution of the variants of *aside* in the corpus

As for the rest of the map, there is a clear fringe of black symbols, which correspond to the conservative, pre-coalescence types *on side/on syde*, in the areas of the North and North-East Midlands dialects (notably Lincolnshire, Yorkshire and Lancashire), where those types are predominantly found. Among these, the type that is most salient is *on syde*, very close in shape to the original OE spelling *on side*, and once more pointing to linguistic conservatism. Finally, there are on the map some noticeable exceptions to the tendencies mentioned below (like the unexpected form *on side* in Gloucestershire), but most of these are related to marginal types, such as the type *azide* in the county of Kent (with just one token, showing the characteristic voicing of the medial /s/) or the form *oside* in the north, in Durham (with just 1 token too). It is also interesting to note that we find the type *on syd* only further north, in Scottish territory.

In conclusion, it seems clear that the northern dialects preserved the use of the more conservative forms of *aside* (notably the two-word, pre-coalescence forms *on side*/*on syde*) to a greater extent than the rest of dialects. This tendency is in sharp contrast with the well-known tendency for the inflectional morphology of the northern dialects of EME to have moved faster in the process of simplification of the OE paradigms than the rest of dialects in Britain (Mossé 1952: 60–61; Bennet and Smithers 1968; Lass 1992: 103ff; Fischer *et al*. 2017: 65 ff). A tentative explanation for this fact might be (as we mentioned before) that the change *on side* > *aside* originated in southern England and spread north thence, following the subsequent process of standardisation of the English language, taking longer to complete in the northern counties.

5. CONCLUDING REMARKS

The results of our analysis allow us to conclude that the grammaticalisation of the adverbial form *aside* in English followed similar paths to other grammatical elements such as connectives or prepositions in the ME period. This implies an increased bondedness between the two former morphemes from which the expression originated in OE (the combination *on side*), and an extension of the semantic range of *side*, moving away from the senses 'part of' or 'in the immediate vicinity to', to expressing an implication of distancing or detachment either in the concrete or abstract sense. There is limited but relevant evidence in the corpus to suggest that the first steps of the grammaticalisation of *aside* were taken perhaps already in the OE period, as one example of the sequence *on side* in that period has the meaning 'off (physically) to one side', which was later extended to cover an array of abstract contexts denoting detachment in the Middle Ages. If this is so, perhaps the chronology of the grammaticalisation of *on side* > *aside* should be reassessed, pushing it back 300 years, into the OE period. There is, notwithstanding, a big gap between the incipient OE grammaticalised use of *aside* mentioned before and the first clear instances of similar meanings in ME quoted in the great historical dictionaries of the English language (ca. 1300). A full understanding of the intermediate steps of the development will probably require a thorough analysis of EME texts (especially from the twelfth century and the first half of the thirteenth century), regrettably a period with a notable dearth of output in English.

A second major conclusion to this work is that the morphosyntactic and phonological processes of change associated with the grammaticalisation of *on side* > *aside* in English (notably, coalescence and attrition) were almost completed in the LME period. This is proven by the presence in the medieval corpus for this work (dated late fourteenth and fifteenth centuries) of a majority of univerbated forms, amounting to 92.5% of the total number of relevant examples, which systematically outnumber two-word ones at every stage of the development, and by the massive presence of reduced forms of the original preposition *on* into *a/o*.

Finally, the geographic distribution of the different variants of *aside* throughout the medieval corpus shows a clear correlation between the most conservative forms (those displaying the pre-coalescence structure with two morphemes and the preservation of the preposition *on*) and a northern origin for the texts. This suggests that the process of grammaticalisation that gave rise to *a*-adverbs in English, and *aside* in particular, might have originated in southern England, spreading north from there.

REFERENCES

Allen, Cynthia. 2019. *Dative External Possessors in Early English*. Oxford: Oxford University Press.

Bennett, Jack Arthur Walter and Geoffrey Victor Smithers. 1968. *Early Middle English Verse and Prose* (second edition). Oxford: Clarendon Press.

Brenda, Maria. 2014. *The Cognitive Perspective on the Polysemy of the English Spatial Preposition* Over. Newcastle upon Tyne: Cambridge Scholars Publishing.

Brunner, Karl. 1913. *Der Mittelenglische Versroman über Richard Löwenherz: Kritische Ausgabe nach allen Handschriften mit Einleitung, Anmerkungen und Deutscher Übersetzung*. Vienna: Wilhelm Braumüller.

Croft, William. 2000. *Explaining Language Change: An Evolutionary Approach*. London: Longman.

Fischer, Olga, Hendrik De Smet and Wim van der Wurff. 2017. *A Brief History of English Syntax*. Cambridge. Cambridge University Press.

Heine, Bernd, Ulrike Claudi and Friederike Hünnemeyer. 1991.*Grammaticalization: A Conceptual Framework.* Chicago: University of Chicago Press.

Heine, Bernd and Tania Kuteva. 2002. *World Lexicon of Grammaticalization*. Cambridge: Cambridge University Press.

Heine, Bernd. 2003. Grammaticalisation. In Brian D. Joseph and Richard D. Janda eds. *The Handbook of Historical Linguistics*. Malden, MA: Blackwell, 575–601.

Hopper, Paul J. and Elizabeth Closs Traugott. 2003. *Grammaticalisation* (second edition). Cambridge: Cambridge University Press.

Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Kemenade, Ans van and Bettelou Los. 2003. Particles and prefixes in Dutch and English. In Geert Booji Geert and Jaap van Marle eds. *Yearbook of Morphology, 2003*. Dordrecht: Kluwer Academic Publishers, 79–117.

Lass, Roger. 1992. Phonology and Morphology. In Norman Blake ed. *The Cambridge History of the English Language. Volume II (1066–1476)*. Cambridge: Cambridge University Press, 23–155.

Lakoff, George and Mark Johnson. 1980. *Metaphors we Live by*. Chicago: The University of Chicago Press.

Lehman, Christian. 1985. Grammaticalization: Synchronic variation and diachronic change. *Lingua e Stile* 20/3: 303–318.

Mossé, Fernand. 1952. *Handbook of Middle English* (ninth edition). Baltimore: Johns Hopkins University Press.

Pérez Lorido, Rodrigo and Patricia Casado Núñez. 2017. Early stages of the '*his* genitive': Separated genitives in Old English. *SELIM, Journal of the Spanish Society for Medieval English Language and Literature* 22: 45–75.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Random House. 1987[1966]. *The Random House Dictionary of the English Language* (second edition, unabridged). New York: Random House.

Rissanen, Matti. 2004. Grammaticalisation from side to side: On the development of *beside(s)*. In Hans Lindquist and Christian Mair eds. *Corpus Approaches to Grammaticalization in English*. Amsterdam: John Benjamins, 151–170.

Schlüter, Julia. 2008. Constraints on the attributive use of 'predicative-only' adjectives: A reassessment. In Graeme Trousdale and Nicholas Gisborne eds. *Constructional Approaches to English Grammar*. Berlin: Mouton de Gruyter, 145–182.

Traugott, Elizabeth Closs. 1999. Subjectification in grammaticalisation. In Dieter Stein and Susan Wright eds. *Subjectivity and Subjectivisation: Linguistic Perspectives*. Cambridge: Cambridge University Press, 32–54.

Tyler, Andrea and Vyvyan Evans. 2003. *The Semantics of English Prepositions*. Cambridge: Cambridge University Press.

DIACHRONIC CORPORA AND HISTORICAL DICTIONARIES

Dictionary of Old English Corpus (DOEC): original release (1981) compiled by Angus Cameron, Ashley Crandell Amos, Sharon Butler and Antonette diPaolo Healey (Toronto: DOE Project 1981); 2009 release compiled by Antonette diPaolo Healey, Joan Holland, Ian McDougall and David McDougall, with TEI-P5 conformant-version by Xin Xiang (Toronto: DOE Project 2009). https://www.doe.utoronto.ca/pages/index.html

Corpus of Middle English Prose and Verse (CMEPV): assembled from works contributed by University of Michigan faculty and from texts provided by the *Oxford Text Archive*, as well as works created specifically for the corpus by the HTI. https://quod.lib.umich.edu/c/cme/

Linguistic Atlas of Late Medieval English (LALME): Michael Benskin, Margaret Laing, Vasilis Karaiskos and Keith Williamson. *An Electronic Version of A Linguistic Atlas of Late Mediaeval English*. http://www.lel.ed.ac.uk/ihd/elalme/elalme.html

Middle English Dictionary. 1952–2001. Hans Kurath, Sherman M. Kuhn and Robert E. Lewis eds. Ann Arbor: University of Michigan Press. Online edition available at the *Middle English Compendium*. 2000–2018. Frances McSparran *et al.* ed. Ann Arbor: University of Michigan Library. https://quod.lib.umich.edu/m/middle-english-dictionary/dictionary

Oxford English Dictionary Online: Oxford University Press. http://www.oed.com

Corresponding author
Rodrigo Pérez Lorido
Departamento de Filología Inglesa, Francesa y Alemana
Campus El Milán
C/ Amparo Pedregal s/n
33011 – Oviedo
e-mail: lorido@uniovi.es

received: March 2020
accepted: July 2020

Appendix: Corpus of Middle English Texts

The following list comprises all the texts that were retrieved from the *CMEPV* and used as the Middle English corpus for this study:

1) *Treatises of fistula in ano: haemorrhoids, and clysters*;
2) *Melusine. Part I.*
3) *Three Middle-English versions of the Rule of St. Benet and two contemporary rituals for the ordination of nuns*.
4) *The right plesaunt and goodly historie of the foure sonnes of Aymon. Englisht from the French by William Caxton, and printed by him about 1489. Ed. from the unique copy, now in the possession of Earl Spencer, with an introduction by Octavia Richardson*.
5) *Caxton's Blanchardyn and Eglantine, c. 1489: from Lord Spencer's unique imperfect copy, completed by the original French and the second English version of 1595*.
6) *Lyf of the noble and Crysten prynce, Charles the Grete*.
7) *The Cambridge ms (University library, Gg. 4.27) of Chaucer's Canterbury tales*;
8) *The Cambridge ms. Dd. 4. 24. of Chaucer's Canterbury tales, completed by the Egerton ms. 2726 (the Haistwell ms) Ed. by Frederick J. Furnivall ...*.
9) *The Ellesmere ms of Chaucer's Canterbury tales*.
10) *The Hengwrt ms of Chaucer's Canterbury tales*.
11) *The Corpus ms (Corpus Christi coll., Oxford) of Chaucer's Canterbury tales. Ed. by Frederick J. Furnivall.*
12) *The Harleian ms. 7334 of Chaucer's Canterbury tales. Ed. by Frederick J. Furnivall.*
13) *The Lansdowne ms of Chaucer's Canterbury tales*.
14) *The Petworth ms. of Chaucer's Canterbury tales. Ed. by Frederick J. Furnivall.*
15) *Geoffrey Chaucer's Troilus and Criseyde*.
16) *The Canterbury tales*.
17) *The "Gest hystoriale" of the destruction of Troy: an alliterative romance tr. from Guido de Colonna's "Hystoria troiana." Now first ed. from the unique ms. in the Hunterian Museum, University of Glasgow, with introduction, notes, and a glossary, by ... Geo. A. Panton, and David Donaldson, esq..*
18) *S. Editha, sive Chronicon vilodunense im Wiltshire dialekt, aus Ms. Cotton. Faustina B III; hrsg. von C. Horstmann*.
19) *Alphabet of tales: an English 15th century translation of the Alphabetum narrationum of Etienne de Besançon, from Additional MS. 25,719 of the British Museum.*
20) *The english register of Godstow nunnery, near Oxford: written about 1450*.
21) *John Gower's Confessio amantis*.
22) *Robert Henryson's The morall fabillis of Esope the Phrygian*.
23) *Robert Henryson's The minor poems of Robert Henryson*.
24) *Hoccleve's works. Ed. by Frederick J. Furnivall.*
25) *Book of the Knight of La Tour-Landry: compiled for the instruction of his daughters: translated from the original French into English in the reign of Henry VI.*

26) *The vision of William concerning Piers the Plowman, together with Vita de Dowel, Dobet, et Dobest, secundum Wit et Resoun, by William Langland (about 1362-1393 A. D.).*
27) *William Langland's The vision of Piers Plowman.*
28) *Mirrour of the blessed lyf of Jesu Christ: a translation of the Latin work entitled Meditationes Vitæ Christi: attributed to Cardinal Bonaventura: made before the Year 1410.*
29) *The pilgrimage of the life of man, English by John Lydgate, A. D. 1426, from the French of Guillaume de Deguileville, A. D. 1330, 1335. The text ed. by F. J. Furnivall ... With introduction, notes, glossary and indexes by Katharine B. Locock ....*
30) *Lydgate's Reson and sensuallyte, ed. from the Fairfax ms. 16 (Bodleian) and the Additional ms. 29, 729 (Brit. mus.) by Ernst Sieper.*
31) *Le Morte Darthur.*
32) *The story of England.*
33) *Works of John Metham: (Amoryus and Cleopes, &c.).*
34) *Dan Michel's Ayenbite of Inwyt: or, Remorse of conscience: Richard Morris's transcription now newly collated with the unique manuscript British Museum MS. Arundel 57, volume 1, text.*
35) *Paston letters and papers of the fifteenth century, Part I.*
36) *Reginald Pecock's Book of faith; a fifteenth century theological tractate, ed. from the ms. in the library of Trinity college, Cambridge, with an introductory essay by J. L. Morison, M. A. ....*
37) *The repressor of over much blaming of the clergy.*
38) *The metrical chronicle of Robert of Gloucester. Edited by William Aldis Wright. Published by the authority of the lords commissioners of Her Majesty's Treasury, under the direction of the master of the rolls.*
39) *Polychronicon Ranulphi Higden maonachi Cestrensis; together with the English translations of John Trevisa and of an unknown writer of the fifteenth century;*
40) *The English works of Wyclif hitherto unprinted. Edited by F. D. Matthew.*
41) *Select English works of John Wyclif; edited from original mss. by Thomas Arnold.*
42) *The Brut, or The chronicles of England. Edited from Ms. Raw. B171, Bodleian Library, &c., by Friedrich W. D. Brie, with introduction, notes, and glossary ....*
43) *An English chronicle of the reigns of Richard II, Henry IV, Henry V, and Henry VI written before the year 1471; with an appendix, containing the 18th and 19th years of Richard II and the Parliament at Bury St. Edmund's, 25th Henry VI and supplementary additions from the Cotton. ms. chronicle called "Eulogium." Edited by John Silvester Davies.*
44) *Political, religious, and love poems. Some by Lydgate, Sir Richard Ros, Henry Baradoun, Wm. Huchen, etc. from the Archbishop of Canterbury's Lambeth Ms. no. 306, and other sources, with a fragment of The Romance of Peare of Provence and the fair Magnelone, and a sketch, with the prolog and epilog, of The Romance of the knight Amoryus and the Lady Cleopes.*
45) *The babees book, Aristotle's A B C, Urbanitatis, Stans puer ad mensam, The lvtille childrenes lvtil boke, The bokes of nurture of Hugh Rhodes and John Russell, Wynkyn de Worde's* Boke *of keruynge, The booke of demeanor, The boke of curtasye, Seager's Schoole of vertue, &c. &c. with some French and latin poems on like subjects, and some forewords on education in early England. Ed. by Frederick J. Furnivall ....*
46) *Hymns to the Virgin & Christ, the Parliament of devils, and other religious poems.*

47) *The book of quinte essence or the fifth being; that is to say, man's heaven. A tretice in Englisch breuely drawe out of Þe book of quintis essencijs in Latyn, Þat Hermys Þe prophete and kyng of Egipt, after Þe flood of Noe fadir of philosophris, hadde by reuelacioun of an aungil of God to him sende. Ed. from the Sloane ms. 73, about 1460-70 A.D., by Frederick J. Furnivall, M.A.*
48) *Early English versions of the Gesta Romanorum*.
49) *Altenglische legenden*.
50) *Companion to the English prose works of Richard Rolle: a selection*.
51) *Sammlung altenglischer legenden, grösstentheils zum ersten male hrsg. von C. Horstmann*.
52) *The romance of Sir Beues of Hamtoun. Ed. from six manuscripts and the old printed copy, with introduction, notes, and glossary, by Eugen Kölbing ....*
53) *Cursor mundi (The cursur o the world). A Northumbrian poem of the XIVth century in four versions. Ed. by the Rev. Richard Morris ....*
54) *Prose life of Alexander*.
55) *Generydes, a romance in seven-line stanzas. Ed. from the unique paper ms. in Trinity college, Cambridge (about 1440 A.D.), by W. Aldis Wright*.
56) *Merlin: or, the early history of King Arthur: a prose romance*.
57) *The Laud Troy book*.
58) *The romance of Guy of Warwick. The second or 15th-century version. Edited from the paper ms. Ff. 2. 38. in the University Library, Cambridge, by Dr. Julius Zupitza ....*
59) *The romance of Guy of Warwick. The first or 14th-century version*.
60) *The Towneley plays*.
61) *The York plays*.
62) *Everyman.*
63) *The Holy Bible, containing the Old and New Testaments, with the Apocryphal books.*

RiCL Research in Corpus Linguistics

Looking into international research groups' digital discursive practices: Criteria and methodological steps in the compilation of the *EUROPRO* digital corpus

Daniel Pascual ─ Pilar Mur-Dueñas ─ Rosa Lorés
University of Zaragoza / Spain

Abstract –The *EUROPRO* digital corpus was designed by the *InterGedi* research group, based at the University of Zaragoza (Spain). The main focus of *InterGedi* is the analysis of the textual resources used by international research groups as part of their dissemination and visibility strategies. The corpus comprises a collection of 30 international research project websites funded by the *European Horizon2020 Programme* (*EUROPROwebs* corpus). By looking into their websites, 20 projects were observed to maintain a *Twitter* account and the tweets from these accounts were the basis for the compilation of the *EUROPROtweets* corpus. This paper delves into the criteria used for the selection of the research project websites and the methodological steps taken to classify, label and tag the verbal component in these websites and tweets. The paper discusses the challenges in the compilation of the corpus because of the dynamic, hypermodal, and hypermedial nature of the digital texts it contains. The paper closes by underlining the potential uses and applications of *EUROPRO* in order to gain insights into the digital discursive and professional practices used by international research groups to foster their visibility online.

Keywords – corpus design; digital discourse; research project websites; *Twitter*; e-visibility; Computer-Mediated Communication

1. INTRODUCTION[1]

Professional practices are increasingly influenced by digital communication. This is no exception for scholars who need to deploy digital discursive practices, especially when it comes to disseminating the results of their research. It is not only necessary for

[1] The compilation of the corpus has been carried out within the framework of the project *Towards Greater Visibility and Dissemination of Scientific Research: A Linguistic, Rhetorical and Pragmatic Study of Digital Genres in English as a Language of International Communication* (grant number FFI2017-84205-P). We would like to thank Dr Enrique Lafuente-Millán for coordinating different stages during the process of compilation. We would also like to express our gratitude to our research colleagues in the *InterGedi* (*Interpersonality in Digital Genres*) research group for their collaboration in this endeavour. Finally, we are also grateful to Dr Rosana Villares Maldonado, a research fellow from our research group *CIRES* (Gobierno de Aragón ─ H16_17R), for providing technical support in the compilation of the corpus.

Research in Corpus Linguistics 8/2: 87–102 (2020). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.08.02.05

academics to produce primary output, which certifies and legitimises new knowledge (Puschmann 2015: 31), but also to disseminate it broadly, which is frequently done online and in English. With the aim of undertaking lexico-grammatical, pragmatic, discursive and genre analyses of digital texts in the international research project websites, the *InterGedi* research group[2] compiled a database of 100 research project websites funded by the *European Horizon2020 Programme*, henceforth H2020. Out of these websites, a corpus of 30 webs fulfilling the criteria described below was compiled (see Section 2.1), and the texts were downloaded and tagged. This digital corpus was named *EUROPRO* and consists of two collections: *EUROPROwebs* which includes the texts downloaded from the 30 research project websites that were selected, and *EUROPROtweets* which includes the tweets from the 20 projects which had a *Twitter* account.

In order to compile the *EUROPRO* digital corpus, the World Wide Web was used 'for' a corpus rather than 'as' a corpus (Fletcher 2013), as a careful selection of webs and their texts was made to compile our own corpus. This is an effective alternative to foster fine-grained analyses but requires solving the challenges of the digital environment before the texts can be processed and worked on, as will be discussed presently (see Section 2.2). *EUROPRO* can also be described as a specialised and static corpus (Gries and Newman 2013: 259), since it delves into the context of international research communication by compiling, at a given point, digital instances of researchers' discursive practices for the dissemination of their projects. The analysis of this type of discourse through specialised and ready-made corpora, as is the case in the *EUROPRO* digital corpus, is advantageous because such corpora tend to be of a manageable size. Compiling a specialised ready-made corpus allows for more qualitative analyses and may help overcome the de-contextualisation of texts and of the particular discursive and linguistic features that are analysed. Such a de-contextualisation of texts and their features is a frequently criticised methodological aspect in corpus-based studies. In this sense, the present paper intends to provide metadata which may contribute to contextualising the digital material included in the corpus, thus facilitating its use and its application.

[2] For information on the research group, see http://intergedi.unizar.es/

The *EUROPRO* digital corpus is also static in that the texts were compiled at a specific moment and not modified or extended, regardless of their evolution in the digital sources where they are hosted. As a result, and to deal with the organic nature of websites and *Twitter* accounts, our corpus captures texts as published final products which constitute the basis for our analyses. The compilation of the corpus over time would have been complex and so would have been its use, since multiple, on-going versions of the texts would have been available. Instead, we decided *EUROPRO* to remain static and to retrieve enough contextual information, as to cater for the process of text crafting and publication.

While it is true that texts online can be easily accessed and saved, important challenges and decisions need to be made when compiling corpora emerging from 'Computer-Mediated Communication', since these are distinct from those conformed by off-line texts or speech genres (Collins 2019). Some of these challenges are related to the selection of the specific sites and the texts in the corpora to ensure representativeness and to the coding of contextual information, which is of great relevance, especially in the website. The *EUROPRO* digital corpus contains textual documents in a reduced form consisting of character strings (Beißwenger and Storrer 2008: 297), but including a prominent layout and structure as well as multimodal elements (see Section 2), in the understanding that the combination of different modes —verbal, visual and audiovisual— makes meaning as a multimodal ensemble (Kress and van Leeuwen 2001; Jewitt 2016).

The aim of this paper is to describe the selection and nature of the texts compiled in the *EUROPRO* digital corpus and to justify the criteria followed in its compilation as regards size, balance, representativeness and topic (Sinclair 2005). Section 2 provides the description of the corpus. Section 2.1 discusses the criteria followed for its compilation while Section 2.2 outlines the methodological aspects considered when compiling, downloading and storing such texts in the belief that texts need to be gathered according to explicit design criteria (Tognini-Bonelli 2001: 2). Section 2.3 offers information on the process of labelling and tagging the *EUROPRO* digital corpus and Section 2.4 describes some important contextual factors of the corpus. Finally, Section 3 deals with the uses and applications of the corpus and Section 4 provides some concluding remarks.

2. Description of the corpus

The websites and *Twitter* accounts selected for the *EUROPRO* digital corpus emerge as part of the communication and dissemination plans included in the *European Horizon2020 Framework Programme for Research and Innovation*. Thus, they are taken to be instances of current digital scientific writing practices which, among others, serve the purpose of accounting for the adequate investment of public expenditure.

From a technical viewpoint, research project websites, as all websites, contain fluid texts which are featured by hypermediality and divided into webpages. These webpages play the role of dynamic nodes and host verbal and audiovisual content that users can navigate through hyperlinks (Djonov 2007: 145). In the case of the research project websites, they render visible a hybrid nature interweaving traditional and digital sources in their design and content organisation. Consequently, the overall function of the genres and texts housed in the research project websites sits "uneasily somewhere between a commercial, technical description of the product and a more formal report on facts" (Stein 2006: 5). These websites serve as repositories of the activities and productivity of the project, as transmitters of the current values of scientific research and as venues to strategically engage with interested users and make the research available to a broad audience (Lorés 2020: 1).

The tweets gathered in the *EUROPROtweets* collection are intended to illustrate the use that research groups make of social networks for the dissemination of their projects, assuming that "interactions in social media contexts may enable self-promotion strategies that result in social or economic gain" (Page 2012: 182). The dynamicity, immediacy and addressivity of social media such as *Twitter* enable research groups to develop a distinct kind of communication about the projects. In *Twitter*, they report on scientific progress and also devote space to daily issues, related topics and social and professional bonds. Hence, tweets are maximised to mediate everyday routines of professional research work and connect users, collaborators and beneficiaries at scholarly events (Kuteeva 2016: 440). In all, *Twitter* users may enact an 'ambient identity' to address the mass online audience and to construe an experience of semiotic belonging to different groups (Zappavigna 2014: 2–3). Such an identity originates in users' discourse choices, in the values conveyed through them and in their exploitation of the affordances of this medium.

2.1. Criteria for the compilation of the EUROPRO *digital corpus*

The *EUROPRO* digital corpus emerges from a database of 100 websites of H2020 projects. It comprises two collections of texts: *EUROPROwebs* and *EUROPROtweets*. The former contains 30 research project websites funded under the H2020 programme with a word count of 394,072 and an average of 13,136 words per website. We here followed Biber *et al.* (1998: 243), who point out that "representativeness refers to the extent to which a sample includes the full range of variability in a population." Therefore, we attempted to collect a specialised corpus which included samples of websites, within the context of the H2020 programme, and would allow to draw reliable insights on aspects of genre theory, metadiscourse, pragmatic strategies and multimodality.

Several criteria were established on the selection of the 30 websites for the corpus to be "a good sampling" (Koester 2010: 69). First, the research projects whose websites were selected had to aim at knowledge creation and dissemination in their respective disciplinary fields and not at training PhD students or professionals. Second, we followed a convenience sampling method which entailed choosing research projects with at least one member from the University of Zaragoza (Spain) or a research institution based in Zaragoza. This allowed to complement our text-based analysis with valuable contextual evidence from potential informants, as discussed in Section 3. Third, at some point, the date of the projects should coincide with the development of our own research project (2018–2021), so that the most recent digital academic practices could be studied.

Given the importance of social media in general —and *Twitter* in particular— for dissemination purposes, *EUROPROtweets* was compiled as an extension of *EUROPROwebs*. This collection of *Twitter* accounts of 20 research projects consists of 4,219 tweets containing 88,970 words, with an average of 211 tweets and 4,449 words per account. The use of social media is indeed highlighted by the communication plans endorsed by Horizon2020 and is generally adopted by research groups. The choice of *Twitter* as the object of study was also based on an observational analysis of the range of social networks maintained by the H2020 research projects within the representative sample, as described in Table 1 below. Out of the 30 research projects, only eight did not make use of any social networks to disseminate their research results. 20 research projects made use of *Twitter*, which was the most frequent social network, 13 made use of *LinkedIn*, ten made use of *Facebook* and five made use of *YouTube*. Interestingly, it was also observed that over 50% of the webs were linked to, at least, two social networks, as shown in Table 1.

NUMBER OF SOCIAL NETWORKS	NUMBER OF H2020 PROJECTS	PERCENTAGE OF H2020 PROJECTS
0	8	26.6%
1	6	20%
2	8	26.6%
3	6	20%
4	2	6.6 %

Table 1: Range of social networks maintained in *EUROPROtweets* by the research projects of H2020

2.2. Methodological decisions in the compilation of the EUROPRO *digital corpus*

Once the selection of websites was determined, several methodological decisions were made for downloading and storing the texts. All texts from the websites were downloaded and labelled using different codes which referred to the pages or sections of the website where such texts were housed, as illustrated in Table 2. We excluded those texts housed in the websites as external downloadable documents, mainly deliverables in PDF format, as they were not considered to share the same digital nature and purpose as those texts generated for the website.

DIFFERENT LABELS USED FOR THE SECTIONS	CODE FOR MENU SECTIONS
Home; Homepage	HOME
About; Objectives; Project; Summary	ABOUT
Partners; Researchers; Consortium; Related projects	PARTNERS
Work packages; Actions; Demos	WORK
News; Events; News and events; Blog	NEWS
Outreach; Publications; Reports; Deliverables; Repository	OUTPUT

Table 2: Codification for salient menu sections and range of headings of sections in *EUROPROwebs*

After that, information related to the extent to which the text could be directly accessed from the website menu was also recorded. In cases where the menu sections were included in the codes provided above, no additional information was required. However, if the section showed up in an unfolding menu of options, the code SUB (subordinate) was noted. Likewise, if the section was included in the website but had no label in the menu, the code EMB (embedded) was added. In this way, a representation of the options of the menu section could give us insights into the preferred position and relevance of some sections or pages throughout the websites.

All texts were downloaded in May 2019 because of the dynamic nature of the research project websites. At that time, the projects in the sample had been developed to different extents just like their websites. For this reason, it was key to record information about the start and end date of the projects, as well as information about their degree of development when the texts were compiled. This information may be of great importance when discussing and interpreting the data retrieved from the analysis. Compiling a 'Monitor Corpus' (McEnery and Wilson 2001), which would need to be updated regularly, was disregarded as a feasible objective in our own research project. Such a compilation would entail constantly comparing updated versions of the websites and tracing them at different points in time throughout the duration of the project which, adding to being extremely time-consuming, would not have been relevant for the purposes of our research.

The 'hypermodal' and 'hypermedial' nature (Petroni 2014) of the websites led to the tagging of (external, internal and peripheral) hyperlinks,[3] of visuals, such as tables, figures, pictures or logos, and of videos and audios. Thus, although the focus of the analyses, as well as the corpus, is mainly grounded on the verbal component, these multimodal and multimedial elements were not overlooked in websites, since they are affordances that combine with the verbal component as meaning-making devices (Kress and van Leeuwen 2001; Jewitt 2016). Similarly, because of the importance of layout and web design aspects, we stored screenshots for every page in the website as part of

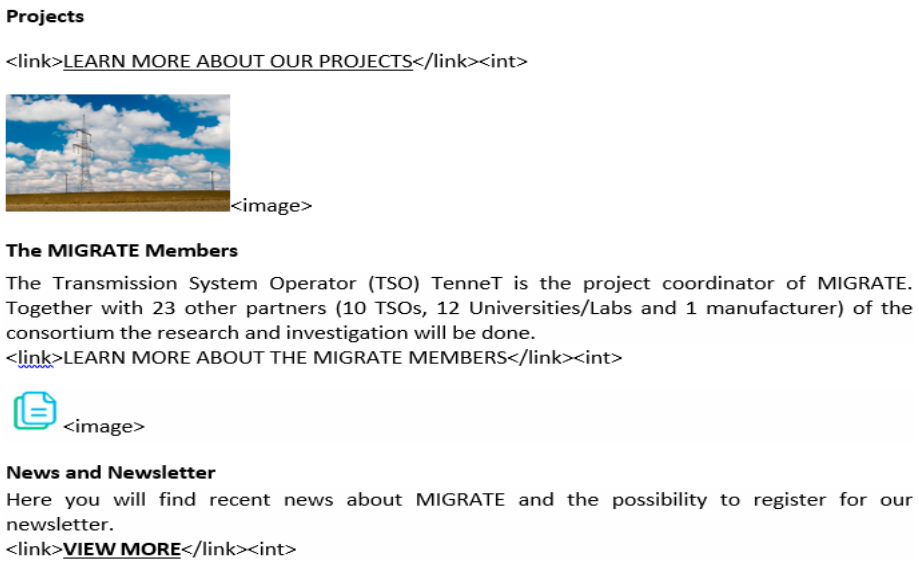
[3] Internal hyperlinks specifically link to other sites within the project website while external hyperlinks lead to external sources of information. Peripheral hyperlinks refer to project-related pages and downloadable documents which are located outside the project website.

the corpus, since this would allow us to go back to them when analysing verbal features in the research project websites.

The methodological decisions to compile *EUROPROtweets* were similar to those taken in the compilation of *EUROPROwebs*. Here, given the dynamic nature of *Twitter* accounts, all tweets and retweets were downloaded in June 2019. Since these platforms are also hypermodal and hypermedial (Petroni 2014), tweets were coded and tagged for hyperlinks and multimodal elements such as pictures, videos or GIFs. One key feature of *Twitter* accounts is their potential interactivity. As a result, we retrieved and saved information about (1) the number of likes in each tweet at the specific date, (2) the number of retweets by other users and (3) the number and types of hashtags (#) used in the tweets by the research group and their mentions (@) to other *Twitter* accounts. This information should be taken into account when carrying out textual analyses based on this corpus, as there may be a correlation between discursive choices and their likely dynamic nature.

2.3. Labelling and tagging of the EUROPRO *digital corpus*

The verbal component of the 30 websites of H2020 which conform *EUROPROwebs* was downloaded and saved into TXT format documents, both as a document corresponding to the whole content of the website (labelled 'NAMEOFTHEPROJECT') and as documents corresponding to common web sections or pages which were labelled with the codes pointed out in Table 2. The tagging of *EUROPROwebs* was performed manually. First, a number of general tags was determined in the light of the texts downloaded and saved, namely <hyperlink>, <image>, <video>, <table>, <graph>, <map>, <presentation>, <questionnaire> and <language>. The whole corpus was then annotated using these tags, as illustrated in Figure 1 below. Specific codes — metadiscursive, pragmatic, ethnographic or multimodal— were further applied to the different analyses undertaken by the members of the *InterGedi* research group.



Projects

<link>LEARN MORE ABOUT OUR PROJECTS</link><int>

<image>

The MIGRATE Members

The Transmission System Operator (TSO) TenneT is the project coordinator of MIGRATE. Together with 23 other partners (10 TSOs, 12 Universities/Labs and 1 manufacturer) of the consortium the research and investigation will be done.
<link>LEARN MORE ABOUT THE MIGRATE MEMBERS</link><int>

<image>

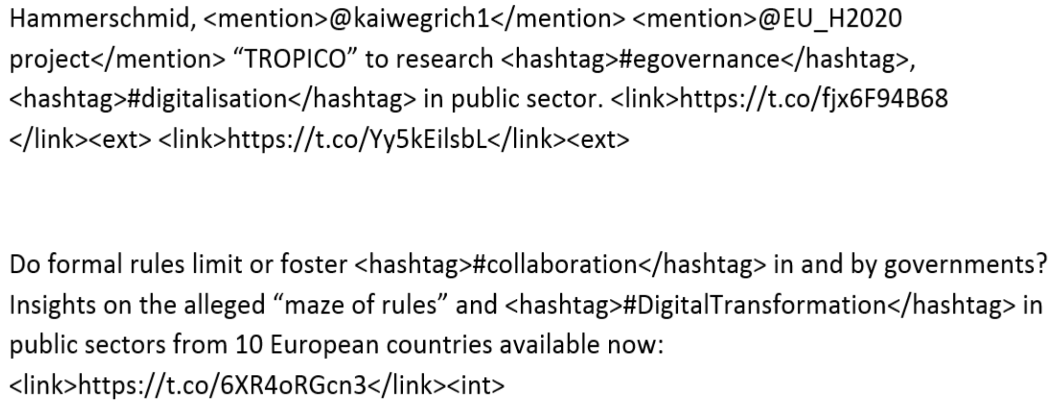
News and Newsletter

Here you will find recent news about MIGRATE and the possibility to register for our newsletter.
<link>**VIEW MORE**</link><int>

Figure 1: Example of manual tagging of *EUROPROwebs*

Manual annotation was leveraged over XML language in the belief that interpretation and evaluation play a significant role in this process at a syntactic, semantic, discursive and pragmatic level (Collins 2019) in the specialised texts that make up the *EUROPRO* digital corpus. Since 'big data' tends to be advocated for in the use of XML —and since the corpus does not contain large collections of dissimilar texts— a consistent coding and tagging carried out by the *InterGedi* members was preferred. This system facilitated keeping in mind the design and layout of the sites, pages and texts under analysis at all times and in a clear visual way.

A similar procedure was followed in the compilation of *EUROPROtweets*. The verbal component of the tweets was downloaded and saved into a TXT document corresponding to each account and labelled 'NAMEOFTHEPROJECT_T'. The tagging of *EUROPROtweets* was also carried out manually. As Figure 2 illustrates, several tags were determined in the light of the texts downloaded and saved: <link>, <hashtag>, <image>, <video> and <language>.



Hammerschmid, <mention>@kaiwegrich1</mention> <mention>@EU_H2020 project</mention> "TROPICO" to research <hashtag>#egovernance</hashtag>, <hashtag>#digitalisation</hashtag> in public sector. <link>https://t.co/fjx6F94B68 </link><ext> <link>https://t.co/Yy5kEilsbL</link><ext>

Do formal rules limit or foster <hashtag>#collaboration</hashtag> in and by governments? Insights on the alleged "maze of rules" and <hashtag>#DigitalTransformation</hashtag> in public sectors from 10 European countries available now: <link>https://t.co/6XR4oRGcn3</link><int>

Figure 2: Example of manual tagging of *EUROPROtweet*s

2.4. Contextual factors of the EUROPRO *digital corpus*

The *EUROPRO* digital corpus was designed to undertake generic, discursive, pragmatic and multimodal analyses of digital texts. In this endeavour, contextual information about the texts to be analysed was essential. For this reason, special efforts were made to obtain and code specific details about the H2020 programme in general, and about each of the research projects selected to compile the *EUROPROwebs* corpus in particular. Table 3 provides an example of the contextual information recorded in the first three websites of our data set. This information mainly includes: (1) the name of the research project, (2) the link to the *CORDIS*[4] web where the details of the project can be accessed, (3) the start and end date of the project and (4) information about the researcher affiliated to the University of Zaragoza or to the research institution based in Zaragoza.

RESEARCH PROJECT	**LINK TO H2020 CORDIS**	**LINK TO WEBSITE**	**PROJECT START AND END DATE**	**DATE OF WEBSITE DOWNLOAD**	**CONTACT AND E-MAIL**
ADREM	cordis.europa.eu/ project/id/680777	spire2030.eu/ adrem	01/10/2015- 30/09/2019	07/05/2019	xxx
AGROinLOG	cordis.europa.eu/ project/id/727961	agroinlog- h2020.eu/en/home/	01/11/2016- 30/04/2020	13/05/2019	xxx
AIDA-2020	cordis.europa.eu/ project/id/654168	aida2020.web. cern.ch/	01/05/2015- 30/04/2019	26/05/2019	xxx

Table 3: Example of contextual information recorded about some of the projects in the *EUROPRO* corpus

[4] *CORDIS* is a platform that gathers information on EU-funded Projects of Research and Development activities. It is the primary source for the consultation of updates about the results and publications of projects participating in different European programs, among with *HorizonH2020* is included. Access is available at https://cordis.europa.eu/.

3. CORPUS APPLICATIONS FOR THE ANALYSIS OF DIGITAL ACADEMIC DISCOURSE

The *EUROPRO* digital corpus is a corpus "that may serve to support empirical research on linguistic aspects of Computer-Mediated Communication discourse" (Beißwenger and Storrer 2008: 293), more specifically on digital academic communication within the context of international research projects. Consequently, its analysis can be undertaken from a range of linguistic theoretical frameworks and analytical perspectives, such as Computer-Mediated Communication, corpus linguistics, discourse analysis, ethnography studies, pragmatics, metadiscourse and multimodality. All of these can help to understand how research groups communicate their results online and allow to delve into other associated concepts such as knowledge dissemination, project accountability and e-visibility.

Studies of a contrastive nature can be carried out using the *EUROPRO* digital corpus for the purpose of looking for features that characterise digital research communication. Such features would be necessarily determined by both the technological level, which deals with the type and degree of exploitation of web-mediated affordances, and the linguistic or discursive level, which focuses on the ways language is employed. Two main directions are reckoned to be particularly rewarding when undertaking contrastive analyses around *EUROPRO*. First, a comparative study of *EUROPROwebs* and *EUROPROtweets* involves contrasting the discourse attested in the websites of the projects with that of *Twitter* accounts held by research groups. Such a study would cast light on the similarities and differences in the use of discursive resources made in both digital platforms when it comes to disseminating and promoting the project and the investigation that is being carried out. Second, findings from the analysis of *EUROPRO* can be contrasted with those from studies on other digital modes and media for research dissemination purposes, such as research reports and research group blogs.

A closer look at the *EUROPROwebs* collection can also be taken at a more rhetorical level by carrying out analyses that focus on the exploration and comparison of sections and/or pages. Here, studies of move analysis (Swales 1990, 2004), which have proved to be insightful for offline academic texts, may help explain potential structural, textual and discursive choices in the different webpages within the research project website, such as 'About', 'Partners', 'Work Packages', 'Output' or 'News and Events'. In turn, move analysis may also allow the identification of prototypical patterns that stand out throughout the corpus and help to generalise researchers' choices in these discursive practices. Furthermore, the relationship between tweets and webs can also be explored from the perspective of genre studies by identifying the genre relations that can be established among texts in both platforms in the light of concepts such as 'generic integrity', 'genre colonies' or 'genre constellations' (Bhatia 2004).

Pragmatic analyses of both *EUROPROwebs* and *EUROPROtweets* may foreground the identification and reasoning of researchers' intents when communicating their projects digitally. The study of the pragmatic mechanisms and resources exploited to disseminate the research that is undertaken can be approached from various pragmatic theories such as Speech Act Theory (Austin 1965; Searle 1969), Relevance Theory (Sperber and Wilson 1995) or Politeness Theory (Brown and Levinson 1987). Pragmatically speaking, the comparison of the aforementioned aspects in the research project websites, as opposed to social networks, such as *Twitter*, may lead to discover researchers' dissimilar intents and strategies that depend on the digital environment employed.

Moreover, the discursive analysis of the *EUROPRO* digital corpus may contribute to exploring the actual use of the language in such a digital scenario. The spectrum of linguistic items deployed in the communication of the project could be accounted for through different types of studies, for instance, at the lexico-grammatical level or at the level of metadiscourse. Additionally, corpus-assisted analyses (e.g. frequency, keywords, collocation and cluster analyses) could make a significant contribution at unveiling meaningful patterns by offering quantitative data. The findings at the discursive level would surely help establish connections with analyses at the rhetorical and pragmatic levels, ranging from a rather abstract and implicit level to the linguistic components that are used.

These analyses of the *EUROPRO* digital corpus —at the rhetorical, pragmatic and discourse levels— should ideally be combined with ethnographically-informed qualitative data. In other words, contextual information gathered from informants in ethnographic analyses can complement and expand the textual results. Thanks to the sampling method followed in the compilation of *EUROPRO* (see Section 2.1), this sort of evidence is at hand. Hence, the role of researchers and their attitude towards the digital communication used in their research projects can be unravelled. This would allow for a better explanation on how international funded research projects develop their research and make their results visible online.

Finally, multimodal analyses are also necessary because there is a need to understand how the combination of languages, modes and media works in those texts hosted in the research project websites and *Twitter* accounts. This may contribute to unveiling the discursive and pragmatic functions that elements such as images, videos, interactive visuals, hyperlinks and other technical affordances perform in digital communication in general, and in Computer-Mediated scientific Communication in particular.

4. SUMMARY AND CONCLUSIONS

As has been pointed out, the *EUROPRO* digital corpus has been compiled to cater for the need to analyse scholarly discourse and scientific communication, as adapted to the digital environment. To narrow down the scope of the various possibilities offered by the Internet to develop such discursive practices, two digital platforms were specifically chosen, namely research project websites and *Twitter* accounts. Accordingly, two collections of texts make up the *EUROPRO* digital corpus: *EUROPROwebs* (texts downloaded from research project websites) and *EUROPROtweets* (texts retrieved from *Twitter* accounts of research projects). Thus, *EUROPRO* will allow to explore different web-mediated affordances and user-dependent linguistic decisions when disseminating research online.

In this paper, the emphasis has been placed on the criteria used to compile the *EUROPRO* digital corpus, which ensure the identification of current digital practices through the exploration of the texts, and the necessary contextual information around them for the analyses to be carried out. Moreover, methodological explanations have been offered to determine how digital texts of an inherent dynamic, fluid, hypermodal and hypermedial nature have been dealt with when compiling a static corpus, reflecting on potential hurdles posed by technical and structural features of the sites where the texts are hosted. Given the scarce number of specialised digital corpora, we believe that the decisions we have made in the compilation of the corpus can help others in the compilation of future corpora.

Furthermore, the fact that the specialised corpus we have compiled may serve a wide range of applications encourages us to make it publicly available in the near future, once the objectives in our current national research project have been achieved. We believe that this corpus may be of use to scholars interested in websites as a digital environment for exploration in genre studies. It will also be of interest to academics who conceive websites and social networks as spaces of engagement for scientific communication and interaction.

Finally, potential uses of the *EUROPRO* digital corpus have also been outlined in the paper. These uses comprise discourse and pragmatic studies, contrastive analyses between texts on websites and on *Twitter* or across website sections, as well as analyses that would complement the textual evidence, either from a multimodal perspective or from ethnographically-collected data. The compilation of *EUROPRO* is the first step to carry out analyses that will allow to gain insights into new, changing, digital discursive and professional practices of researchers nowadays.

References

Austin, John L. 1965. *How to Do Things with Words.* Oxford: Oxford University Press.

Beißwenger, Michael and Angelika Storrer. 2008. Corpora of computer-mediated communication. In Anke Lüdeling and Merja Kytö eds. *Corpus Linguistics: An International Handbook*. Berlin: Mouton de Gruyter, 292–308.

Bhatia, Vijay K. 2004. *Worlds of Written Discourse: A Genre-based View*. London: Continuum.

Biber, Douglas, Susan Conrad and Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Brown, Penelope and Stephen Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Collins, Luke Curtis. 2019. *Corpus Linguistics for Online Communication: A Guide for Research*. London: Routledge.

Djonov, Emilia. 2007. Website hierarchy and the interaction between content organization, webpage and navigation design. *Information Design Journal* 15/2: 144–162.

Fletcher, William H. 2013. Corpus analysis of the World Wide Web. In Carol Chapelle ed. *Encyclopedia of Applied Linguistics: Volume 3.* New Jersey: Wiley-Blackwell, 1339–1347.

Gries, Stefan T. and John Newman. 2013. Creating and using corpora. In Robert J. Podesva and Devyani Sharma eds. *Research Methods in Linguistics*. Cambridge: Cambridge University Press, 257–287.

Jewitt, Carey. 2016. Multimodal analysis. In Alexandra Georgakopoulou and Tereza Spilioti eds. *Routledge Handbook of Language and Digital Communication*. London: Routledge, 69–84.

Koester, Almut. 2010. Building small specialised corpora. In Anee O'Keeffe and Michael McCarthy eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 66–79.

Kress, Gunther and Theo van Leeuwen. 2001. *Multimodal Discourse: The Modes and Media of Contemporary Communication*. London: Arnold.

Kuteeva, Maria. 2016. Research blogs, wikis and tweets. In Ken Hyland and Philip Shaw eds. *The Routledge Handbook of English for Academic Purposes*. London: Routledge, 431–444.

Lorés, Rosa. 2020. Science on the web: The exploration of research websites of energy-related projects as digital genres for the promotion of values. *Discourse, Context and Media* 35: 1–10.

McEnery, Tony and Andrew Wilson. 2001. *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

Page, Ruth. 2012. The linguistics of self-branding and micro-celebrity in Twitter: The role of hashtags. *Discourse & Communication* 6/2: 181–201.

Petroni, Sandra. 2014. Collaborative writing and linking: When technology interacts with genres in meaning construction. In Paola E. Allori, John Bateman and Vijay K. Bhatia eds. *Evolution in Genre*: *Emergence, Variation, Multimodality*. Bern: Peter Lang, 289–306.

Puschmann, Cornelius. 2015. A digital mob in the ivory tower? Context collapse in scholarly communication online. In Marina Bondi, Silvia Cacchiani and Davide Mazzi eds. *Discourse in and through the Media: Recontextualizing and Reconceptualizing Expert Discourse*. Newcastle upon Tyne: Cambridge Scholars Publishing, 22–45.

Searle, John Rogers. 1969. *Speech Acts*: *An Essay in the Philosophy of Language.* Cambridge: Cambridge University Press.

Sinclair, John. 2005. Corpus and text – Basic principles. In Martin Wynne ed. *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books, 1–16.

Sperber, Dan and Deirdre Wilson. 1995. *Relevance: Communication and Cognition*. Oxford: Blackwell.

Stein, Dieter. 2006. The web as a domain-specific genre. *Language@Internet* 3: https://www.languageatinternet.org/articles/2006/374 (10 May, 2020.)

Swales, John. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Swales, John. 2004. *Research Genres: Explorations and Applications*. Cambridge: Cambridge University Press.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Zappavigna, Michele. 2014. Enacting identity in microblogging through ambient affiliation. *Discourse & Communication* 8/2*:* 209–228

Corresponding author
Daniel Pascual
University of Zaragoza
Department of English and German Studies
Calle San Juan Bosco, 7
50009. Zaragoza
Spain
e-mail: dpascual@unizar.es

received: January 2020
accepted: July 2020

RiCL Research in Corpus Linguistics

A corpus-assisted genre analysis of the *Tunisian Lecture Corpus*: An exploratory study

Basma Bouziri
University of Gabés / Tunisia

Abstract –Multimodal, specialized corpora of academic lectures represent authentic classroom data that practitioners can draw on to design academic listening resources that would help students attend lectures. These corpora can also act as reflective practice corpora for teacher training or professional development programs with the objective of raising awareness of lecturing practices. Despite their contribution in shaping the type and quality of the learning that takes place in classrooms, multimodal lecture corpora are scarce, particularly in the Arab world. This paper addresses this research gap by designing and collecting a corpus of academic lectures delivered in English in Tunisia. The corpus was explored using a Systemic Functional Linguistics and English for Specific Purposes integrated genre analysis framework. A three-layered model of analysis was used to manually code various rhetorical functions as well as their realizations. Major findings include the pervasiveness of metadiscursive functions when compared to discourse functions, the identification of context-specific metadiscursive strategies, and the absence of verbal or non-verbal signaling of some rhetorical functions. Implications relate to the necessity of compiling and/or using lecture corpora that are multimodal, the value of adopting function-first approaches to explore these, particularly in non-native contexts, and the design of professional development programs and learning materials that would better account for local academic needs.

Keywords – Academic lectures; corpus; genre analysis; Systemic Functional Linguistics; English for Specific Purposes; Tunisia; exploratory studies

1. Rationale

In Tunisia, increasing attention and efforts are being devoted to quality pedagogy and teacher training in higher education. To contribute to these efforts, the study of the rhetorical features of lecturers' discourse and the way they are realized is a necessary step. Such research would lead to designing needs and context-specific courses and materials that would support students when attending lectures. It would also play a role in the design of teacher professional development programs with the aim of upgrading the quality of teaching in higher education. To study lectures, the use of specialized

Research in Corpus Linguistics 8/2: 103–132 (2020). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.08.02.06

corpora is pivotal. Nesi (2008: 1–2) maintains that "as teachers of languages for specific purposes, it is these small specialized corpora that interest us most." The main reason is that the analyst can integrate macro-contextual features that are essential for a sound interpretation of the corpus data (Flowerdew 2004; Camiciottoli 2008). In doing so, explanatory adequacy supplements the descriptive power that corpora already have (Bhatia 2002). The validity of the analysis and interpretation of the corpus data within its context of use is further accentuated when the compilers of specialized corpora are themselves the analysts.

In spite of the value that locally designed and specialized corpora have, they remain scarce. In the Tunisian context, some research has been conducted on classroom discourse (Abdesslem 1987; Touati 2004). However, to our knowledge, there is no study which has been carried out on academic lecture discourse in Tunisia except for Bouziri (2019). One major reason is that compiling spoken data is a daunting and expensive task, which is coupled with the sensitivity of the data under focus as well as with the current requirements for multimodal data. In fact, the use of muted spoken corpora (Ballier and Martin 2015), that is, transcripts of spoken language which are not distributed with audio and video files, does not account for the most basic and immediate context behind their production and their use fails to deliver a valid analysis and interpretation of the data (Deroey and Taverniers 2011). A second reason is the nature of the analyses that are conducted on academic lectures. Since pedagogical applications often represent their ultimate aim, genre analyses of lectures are particularly valuable. In focusing on macro discourse structures and functions, genre approaches draw one of the most significant links between corpora and contexts (Partington 2004) reflecting principled variations not often captured by large-scale corpus studies, since their interest is in the texts and contexts that generated the corpus data (Flowerdew 2013). Additionally, corpus-assisted genre analyses hold the potential for accounting for less marked discourse phenomena.

Although some macro approaches to lecture discourse analysis have been conducted in different contexts (cf. Young 1994; Alsop and Nesi 2014; Bouziri 2019), they are not as widespread as more form-based types of analyses. Along with data transcription, the manual annotation of lectures is, in fact, time-consuming and cognitively demanding because multiple viewings of the lectures and readings of the transcripts are necessary to identify and describe various rhetorical categories. These difficulties are heightened due to the idiosyncratic nature of the lecture genre. In fact, when lecturing, “the lecturer is not under great pressure to exhibit control over conventionalized rhetorical structure” (Thompson 1994: 182). This is unlike the research article genre, for example, where researchers are pressured into adhering to international standards of writing for their work to be published in international journals (Abdesslem and Costello 2018). One reason for the idiosyncratic nature of lectures is that they are live events. This necessarily involves a certain degree of spontaneity and, hence, unpredictability. Another reason is the lecturers’ individual lecturing styles which also account for the high variability that often characterizes this genre along with class size (Lee 2009; Cheng 2012), discipline (Young 1990; Thompson 1994; Deroey and Taverniers 2011), and culture (Alsop and Nesi 2014). All these variables sustain the rhetorical variation that academic lectures exhibit and thus makes their study challenging.

2. The study

To address the aforementioned research gaps and challenges, the *Tunisian Lecture Corpus* (TLC) project was started up. In this paper, I report on and discuss its collection, transcription, and coding with the objective of providing tools for the study of lectures in under-investigated contexts.[1] To this aim, I propose a theoretical framework and a model of genre analysis that are compatible with the specificity of the academic lecture genre. The framework and the model were used to approach the corpus and develop the coding scheme employed for the manual annotation of the various rhetorical functions in TLC. Accordingly, the subsequent sections are organized as follows. Section 3 presents and discusses the theoretical framework of the study: an integrated genre analysis framework that draws on both the Systemic Functional Linguistics (SFL) and English for Specific Purposes (ESP) traditions. In this section, the model of analysis is also presented. Section 4 provides a description of TLC with information about its collection, the participants, the corpus transcription, and its coding and analysis. The results of the corpus-driven study are then presented in Section 5. Section 6 summarizes the findings and discusses their implications.

[1] The transcription conventions and coding scheme developed and used in this study are available in the IRIS database following this link: https://www.iris-database.org/iris/app/home/detail?id=york:938327

3. Theoretical framework

Genre Analysis (GA) is a discourse approach that concentrates on both the linguistic and contextual aspects of texts in specific genres. The analysis is 'top-down' (Biber *et al.* 2007) starting from the rhetorical functions and moving down to their linguistic, non-linguistic, and/or multimodal realization. Within GA, three research traditions have been distinguished: New Rhetoric (NR), Systemic Functional Linguistics (SFL), and the English for Specific Purposes (ESP) approach. In this paper, I will concentrate on the SFL and ESP approaches for two reasons. The first is that NR adopts a social rather than a pedagogical orientation, considering the classroom as "an inauthentic environment lacking the conditions for complex negotiation and multiple audiences" (Hyland 2002: 114). The second is that New Rhetoricians adopt a non-linguistic approach to GA (Hyon 1996; Flowerdew 2002), investigating texts from an ethnographic rather than a discourse analytic perspective.

As opposed to New Rhetoricians, SFL and ESP genre analysts conduct both functional and linguistic analyses of genres with pedagogical motives in mind. Within these two approaches, functional categories are set as the starting point for the analysis. Subsequently, the linguistic features that characterize them are described. According to Callies (2015), such function-driven approaches are rarely implemented in linguistic research despite the fact that they are valuable particularly for research that is conducted in non-native contexts. One reason is that function-driven approaches enhance our understanding of the way forms correlate with the functions they realize in discourse (Callies 2015). They also seek to uncover 'non-canonical' strategies which non-native users may employ to convey meaning, and which can be easily neglected in form-driven approaches (Callies 2015: 54). For the aforementioned reasons, a GA approach has been adopted to analyze TLC.

3.1. Key notions in genre analysis

In this study, three key notions in GA are used. They are drawn from the ESP and SFL approaches. In combination, they are viewed as complementary and relevant to the analysis of the academic lecture genre. The first two constructs identified in the ESP approach are 'moves' and 'steps'. According to Swales (2004: 228–229), a move is:

> a discoursal or rhetorical unit that performs a coherent communicative function in a written or spoken discourse. […] It is a functional, not a formal, unit. […] Sometimes, however, grammatical features can indicate the type or nature of move.

Thus, a move is defined in terms of the rhetorical goal that it seeks to achieve. It is usually "realized in stages […], all of which are more or less steps to the fulfillment of the function of the move" (Bhatia 2001: 86). A step, in turn, constitutes a specific function within a move and serves its higher purpose. Steps have been referred to elsewhere as 'strategies' (Kwan 2006; Yin 2016) or 'sub-functions' (Thompson 1994) to denote their non-obligatory, cyclical, and non-sequential nature (Lee 2009), in contrast to the way they have been interpreted in Swales' model. With respect to academic lectures, for example, several researchers have analyzed introductions as a sub-genre in terms of moves and steps (Lee 2009) or functions and sub-functions (Thompson 1994; Yaakob 2013). They have found that there are multifunctional units and point out that, at times, there is some difficulty to disentangle functions from one another. Another major finding is the lack of "robust preferred orders" (Swales 1990: 145). Thompson (1994), for instance, found that various functions in lecture introductions display a non-sequential structure. Duszak (1994), in her study of academic Polish texts, also reports that moves behave in a cyclical rather than a linear fashion and that various combinations of moves and steps are possible.

The third construct used in the study is that of 'phase' and is drawn from the SFL approach. A phase is defined as a "strand of discourse that recurs discontinuously throughout a particular language event" (Young 1994: 165). Two types of phases are identified: discourse phases and metadiscourse phases. A discourse phase embodies rhetorical functions such as defining, explaining, and exemplifying, whereas metadiscourse phases have structuring, evaluating, and closing functions. Phase boundaries are identified pragmatically rather than temporally using semantic, verbal, non-verbal, and contextual cues. The construct of phase is very useful because it enables the analysis to be conducted via a unit that does not imply any sequential ordering. Phasal analysis is indeed the most influential model proposed to analyze academic lectures within the SFL tradition (Young 1990; Gregory 2002; Wu 2013). As Young (1990: 45) points out, "it describes what actually happens in an instance of discourse where different strands recur and are interwoven to form the discourse plot of an instantiation of language." This kind of analysis is flexible because it does not enforce a linear or hierarchic structure to texts and captures the dynamic and non-sequential nature of the academic lecture discourse. Most importantly, Young's work contributed to a systematic and functionally-oriented analysis of discourse segments where a function rather than a form-based analysis is adopted to identify phases. Another advantage of phasal analysis is that it provides a thick and comprehensive analysis of discourse along both macro-elements and micro-elements.

3.2. An SFL-ESP integrated framework

In combining ESP and SFL approaches to GA within a single theoretical framework, a function-driven model is developed. Phasal analysis is viewed as the analytical approach that best captures the dynamic nature of the academic lecture genre as the construct of phase is soft, non-codal, and non-predictive (Gregory 2002). It is soft, since one phase is likely to occur at any point in discourse and is not restricted to a particular temporal unit such as beginning, middle, or end. In fact, "there are many beginnings, many middles, and many ends." (Young 1994: 165). Similarly, phases are non-codal, as they are recognized and labelled in terms of their function (e.g., evaluation, content, and discourse structuring) and not in terms of an implied hierarchical structure (e.g., initiation, response, feedback). Finally, a phase is not predictive because it does not (necessarily) require a particular phase to occur next.

The terms moves and steps are interpreted in this study as functions and sub-functions, for this conceptualization does not suggest or, at least, impose any kind of a hierarchical order. The two constructs represent more fine-grained rhetorical distinctions that are not captured by phasal analysis. Along with phases, functions and sub-functions would be allowed to occur discontinuously and are characterized by recursiveness and a certain degree of unpredictability.

This looser view of genre is taken up in the SFL-ESP integrated framework that I am proposing in this paper. Within this framework, two interrelated theories and methodologies of genre, SFL and ESP, are assimilated within one single model that draws on their mutual strengths. The theoretical basis and the main components of this framework are summarized as follows:

1. Phases, moves, and steps are rhetorical spaces for the enactment of a set of rhetorical purposes.
2. Phases, moves, and steps are non-predictive and are organized along certain preferences or choices. This means that a significant degree of variability in the exploitation of those rhetorical spaces in the lecture genre should be acknowledged.
3. A fine-grained coding of rhetorical functions is compatible with a phasal analysis.

The constructs of phase, move, and step are used to design a corpus-assisted genre analysis model to analyze TLC. This model is described in the following section.

3.3. Model of analysis

The present corpus-assisted model aims to explore TLC in view of identifying a niche that would lead to develop a corpus-based study. This research process is depicted in Figure 1 where the first parse in the analysis is corpus-driven, involving "the inspection of corpus evidence" (Tognini-Bonelli 2001: 84).

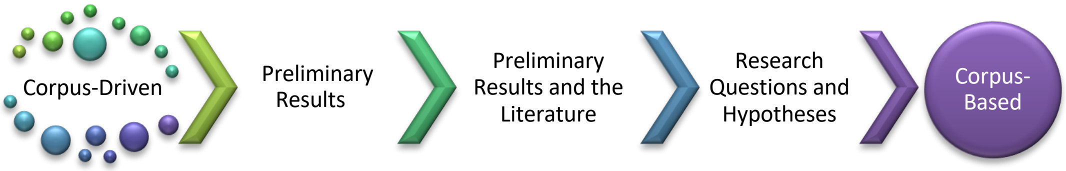

Corpus-Driven
Preliminary Results
Preliminary Results and the Literature
Research Questions and Hypotheses
Corpus-Based

Figure 1: Corpus-driven to corpus-based analyses

The corpus-driven genre analysis set as its first objective the unveiling of rhetorical features that would not be captured by large scale and purely corpus-based techniques (Yaakob 2013). The relationship between phases, moves, and steps in the model is illustrated in Figure 2.

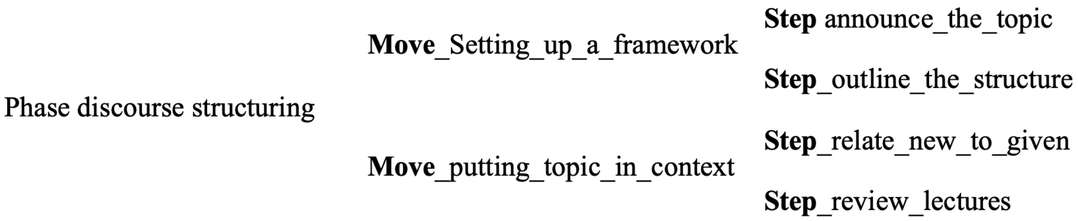

Phase discourse structuring
Move_Setting_up_a_framework
Step announce_the_topic
Step_outline_the_structure
Move_putting_topic_in_context
Step_relate_new_to_given
Step_review_lectures

Figure 2: Rhetorical relations in the *Tunisian Lecture Corpus*

A phase is the upper level category with a general rhetorical function such as structuring, describing, and evaluating. It encompasses more specific moves which carry out its general purpose. Within these rhetorical functions, various steps are identified. This is illustrated in Figure 2 where the discourse structuring phase unfolds into two types of moves: setting up a framework and putting topic in context. In turn, these are divided into two steps each. It should be noted that a discourse structuring phase can be enacted at two different rhetorical planes which I shall call here domains: the content domain and the lecture domain. For example, setting up lecture framework refers to structuring the lecture as a whole, whereas setting up content framework refers to structuring a particular content unit or topic within the content phase.

Regarding their realization, the three rhetorical functions may take different forms: a phrase, a clause, or an utterance-(s). Their boundaries are set based on a mixed set of criteria (Swales 2004). The first and main criterion is linguistic where the meaning of words and expressions such as *I mentioned last week*, *today I'm going to* is used to assign a particular rhetorical function to a segment or to help identify boundaries between two units. A second criterion is the use of paralinguistic features (e.g., intonation, stress), and non-verbal features (e.g., gestures), which can help assign or confirm a particular rhetorical function to a segment. Contextual elements may also be employed to further check and resolve any ambiguity, thus representing the third criterion used during the coding process. Contextual clues can be found in the videos of the lectures, for example. Finally, the coder's knowledge of the way texts within a particular genre and discourse community tend to be structured has also an impact on the coding decisions as well (Dudley-Evans 1994).

4. CORPUS AND METHOD

In this section, details about the corpus and the method used for coding and analyzing it are provided.

4.1. Corpus design

TLC is a non-native, specialized, and multimodal corpus of academic lectures comprising over 106,000 words. It is made up of 12 video recordings and one audio recording. Details of the corpus are shown in Table 1 below.

Words in the corpus	106,200
Mean	8.169
Range	7.913
Class size	Average of 20 students
Hours of recording	20 hours and 50 minutes
Video files	12
Audio files	1
Course description	9 out of 12

Table 1: Overview of the *Tunisian Lecture Corpus*

The recorded lectures took place in two institutions of higher education in Tunisia, with lecturers teaching content courses in English in applied linguistics, cultural studies, and literature. Course descriptions were collected when available, as they can enhance the interpretation of the lectures. The names of participants were kept anonymous, and so were the institutions they were affiliated with. 12 participants provided written consent, and one gave oral consent, to make (some of) their data available. For instance, some participants agreed to make all their data publicly available whereas others agreed only for the transcripts to be shared publicly. All participants were Tunisian non-native speakers of English with Arabic as their native language. Their students were also non-native English speakers and were mostly Tunisians. English was used mainly to teach English language undergraduate and graduate degrees. English-mediated instruction is carried out in some public and private universities for subjects such as business and engineering.

The higher education landscape in Tunisia has also witnessed a shift towards more interactivity in academic lectures. In this regard, several courses now combine the lecture format with the seminar format rather than use the traditional monologic lecture format. The lecture format is adopted to teach theoretical content whereas the seminar format provides a space where students can apply that theory to tasks such as text analysis, oral presentations, or linguistics exercises. Some of the data collected in this study thus reflects a hybrid genre that might not be smilar to data collected in other contexts.

4.2. Data collection

The data was collected during the academic year 2014–2015. Lecturers of undergraduate content courses in four institutions of higher education were contacted. Clarifications and details concerning ambiguous points as well as technical aspects of the recording (e.g., the placement of the video camera and the setting up of the microphone) were discussed. Only 13 lecturers from two institutions granted consent. Recording sessions were then arranged with each one of them during their regular classroom hours. The equipment was tested in a real classroom situation in order to evaluate the quality of the image and sound, comfort in wearing the microphone, the degree of intrusiveness of the mounted camera in class, the students' reactions, and the actual battery and recording capacities. Participants were then recorded for two lectures in a row. One reason was to examine how two different lectures were connected. A second reason was that the first lecture of each participant could provide useful information for the interpretation of the second one which was under investigation. The video camera was mounted on a tripod at the back of the class and angled in order to capture the full front frame including the lecturer, the board, and the whole class with students sitting with their backs to the camera. A background information sheet on each lecture session was also filled. Variables gathered included gender, teaching experience, and language background.

4.3. Data transcription

The transcription system devised is described in Appendix 1. A low-level transcription requirement was opted for and as many relevant contextual elements as possible were included. To balance quality and speed, *Soundscriber*, a transcription tool, was used to walk through the files. To operate this tool, audio files in the wave format were extracted from the video files. In addition to *Soundscriber*, two other windows were used during actual transcription. The first is the video file which was used to check pauses, contextual events, and the second is a plain text file used for transcription. The triangulation of data sources was effective for a smooth and reliable transcription. The transcription process also underwent three passes. Pass one concentrated on textual and specific spoken features like pauses, fillers, and backchannels. Pass two was carried out with the support of video files and was proven essential for a *bona fide* transcription. Indeed, features such as pauses, turn boundaries, and contextual events could be spotted and/or interpreted more appropriately thanks to visual cues. In the final pass, transcripts were edited for more consistency and transcription mistakes were corrected. It should be noted that grammatical mistakes that lecturers made were kept as they were with corrections added via the tag <error corr>. Disfluencies and mistakes pertaining to clausal and utterance structures were not corrected as this would have changed the data.

One last point pertaining to the transcription of students' turns is worth mentioning. Although the focus in this study is on lecturer discourse, students' turns were fully transcribed whenever possible. These were marked between square brackets in the transcripts because, most of the time, they could not be fully and/or clearly heard to be transcribed in any reliable way, and thus fully exploited for the purposes of the current study. In those cases, the speech act performed by the student was transcribed. For example, if a student responded to a question, the response was transcribed as <response>. If a student asked a question, it was transcribed as <question> and so on. Because of the gaps in students' contributions, these were not included in the total number of words in TLC.

4.4. Coding and analysis

To explore TLC, a preliminary coding scheme was devised based on the literature on lecture genre analysis, and thus included some rhetorical functions derived from it. The *UAM CorpusTool* (O'Donnell 2017) was used to draw the scheme and made it possible to manually code the lectures in terms of (pre-)designed features. As the coding proceeded, new categories emerged and were added to the original scheme. At the end of the process, an upgraded version was generated (see Appendix 2). The scheme makes some slight, yet important distinctions between some rhetorical categories. For example, dividing a global rhetorical function such as the discourse structuring phase in terms of more specific rhetorical functions as illustrated in Figure 2 (see Section 3.3) highlights the distinction between two types of knowledge (respectively schematic knowledge and contextual knowledge) that the two moves, (*viz.*, setting up a framework and putting topic in context) activate.

The coding procedure was constructed on a small sub-set of the corpus comprising two lectures (<Civ-09-02-A> and <Ling-07-02-B>), which in turn constituted 10% of the corpus. The procedure involved various stages adapted from Biber *et al.*'s (2007: 12–13) model of 'top-down' approaches to corpus analyses. The first stage included a survey of rhetorical functions. Lecture introductions and lecture closings have been the subject of many research projects and, as such, they represented a major input for a number of rhetorical functions that could be initially included in the coding scheme. As for the rhetorical functions in the content phase, many were drawn from lecture research as well as research on other academic genres. The second stage is the warm-up stage in which the video of the lecture to be coded was viewed. This enabled the coder to get a general feel of the lecture, which in turn guided its interpretation. Notes were taken on details which were not fully captured during transcription, but which were thought to aid the coding process. The video of a previously recorded lecture and the course description were also consulted whenever the coder wanted to obtain further contextual information. The third stage of the coding procedure involved segmenting the text into phases, moves, and steps. Phase and move boundaries can sometimes be fuzzy and, therefore, needed to be constantly revised throughout the coding process in order to reach the finest delimitation of these different rhetorical categories. In the fourth stage, the coding scheme was upgraded and the coder then moved to the analysis of pervasive functions in the corpus in view of pinpointing interesting rhetorical and/or linguistic phenomena. Analysis of the coded categories was carried out using the statistics feature of the *UAM CorpusTool*, which enabled to calculate their frequencies. A qualitative analysis of the coded segments was also conducted in order to study the way the various rhetorical categories were realized verbally and non-verbally.

5. RESULTS AND DISCUSSION

The results of the corpus-driven study are organized into four parts. The first provides general findings about the academic lecture genre and the various functions identified in TLC as observed in the corpus-driven study. The remaining parts focus on the metadiscursive functions identified: their pervasiveness, context-specificity, and verbal and non-verbal realizations.

5.1. Overview

The corpus-driven study confirmed the discontinuous and recurrent nature of phases and moves which were already reported in the literature as a key characteristic of the academic lecture genre. The evaluation phase typically illustrates this discontinuity and recursiveness. Additionally, not all functions weighed equally. A case in point is the difference between the functions: summarizing main points and indicating end in the lecture closing phase. A lecture which closes with a formal indication of an end and another which wraps up content bringing together the main and important points discussed do not have the same pedagogical value. Examining lecture closings, Cheng (2012) also found that reviewing key points has the lowest frequency when compared to the other functions in lecture closings. The finding above was possible thanks to the adoption of the SFL-ESP integrated framework which allowed for the coding of fine-grained rhetorical functions (*viz.*, moves and steps) in addition to more general ones (*viz.*, phases). The framework, particularly its integration of the construct of phase, was also useful as it distinguished between discourse phases and metadiscourse phases.

5.2. Pervasiveness of metadiscursive functions

The present corpus-assisted genre analysis led to identifying a number of rhetorical functions as displayed in Table 2. In Appendix 3, examples of each of the functions listed below are also provided.

RETHORICAL FUNCTION	<Civ-09-02-A>	<Ling-07-02-B>
STRUCTURING	**71**	**98**
FRAMING	**32**	**73**
SETTING_UP_LECTURE_FRAMEWORK		
Announce_the_topic	1	3
Outline_the_structure	0	0
Present_aims	0	0
Announce_start_of_lecture	0	0
Looking_ahead	0	0
SETTING_UP_CONTENT_FRAMEWORK		
Announce_content	19	35
Looking_ahead	2	1
REVIEW _CONTENT		
Indicate_end	1	1
Sum_up	5	32
REVIEW_ LECTURE		
Indicate_end	1	0
Housekeeping	1	0
Looking ahead	1	1
CONTEXTUALIZING	**39**	**25**
PUT_LECTURE TOPIC_IN_CONTEXT		
Refer_to_earlier_lectures	6	0
PUT_CONTENT_IN_CONTEXT		
Refer_to_earlier_content	14	6
Provide_rationale/context	19	19
ELABORATING	**44**	**47**
Explain_content	21	19
Exemplify_content	14	14
Specify_content	2	1
Draw_implication	7	13
DEFINING AND DESCRIBING	**51**	**66**
Define_content	16	14
Describe_content	35	52
EVALUATING_ LECTURE_ CONTENT	**1**	**7**
Show_importance_of_content	0	6
Indicate_Scope	1	1
EVALUATING_KNOWLEDGE	**51**	**19**
Enquire_about_ knowledge	10	5
Give_clues	1	1
Establish_knowledge	12	6
Indicate_attitude	5	0
Indicate certainty/uncertainty	3	1
Give_feedback	16	5
Enhance_understanding	3	1

Table 2: Significant rhetorical functions in two lectures in the *Tunisian Lecture Corpus*

A major observation is the frequency of metadiscursive functions when compared to discourse functions. The structuring function, where the lecturers both frame and contextualize their talk, is the most frequent one in the two lectures under study forming respectively 71 and 99 units. Framing and contextualizing enhance comprehension through the organization of content and the presentation of contextual information that is critical to understand the content at hand. In <Civ-09-02-A>, 44 and 71 units were coded respectively as elaborating and structuring whereas 51 units were coded as defining and describing content. Similarly, <Ling-07-02-B> contains 98 units where the lecturer structures the talk and 47 units where he elaborates on the content introduced. In turn, defining content was coded 66 times. Based on the figures above, it seems that structuring is a key metadiscourse function in TLC. Looking closely at this finding, one can notice that there is a limited amount of lecture framing as reflected in the steps of setting up the lecture framework and reviewing lecture[2] when compared to content framing. This finding may indicate that the structure of lectures may not be obvious for students in order for them to recover the way content is organized in terms of macro and micro points. The pervasiveness of structuring denotes the lecturer's efforts in framing and contextualizing the talk with the aim of achieving coherence between different content units. However, lecture framing is mostly restricted to announcing the topic and indicating end rather than to presenting aims and/or outlining structure. The latter are typically realized in the lecture discourse structuring phases and lecture closing phases. Given the real time conditions under which lectures are delivered, they support and enhance comprehension. It is clear however that these were not fully taken advantage of as rhetorical spaces to carry out important pedagogical functions. In this regard, Palmer-Silveira (2004: 101) states:

> if the introduction is poorly prepared, the audience may lose interest and this can jeopardize the way our students will understand the topic. In the introduction, the audience will need to know the main topic, the purpose, the main concepts we will deal with.

While it is true that content framing contributes to the organization of various points within the lecture, it is nonetheless important for students to have global frames of reference they can resort to during the lecture in order to organize the various information presented. The low frequency of lecture framing has also been reported in the literature on global macro-markers (Thompson 2003; Palmer-Silveira 2004).

Besides structuring, evaluation is a pervasive metadiscursive function. One way in which evaluation unfolds in TLC is through the use of contextual comments as reflected in the function give feedback. Their use was particularly noted in the two lectures under investigation. Contextual comments broadly correspond to commentaries in Vande

[2] Four units each for <Civ-09-02-A> and <Ling-07-02-B>.

Kopple's (1985) and Crismore *et al*'s. (1993) models of metadiscourse, to contextual metadiscourse in Luukka's (1994) and Ädel's (2006) models, and to text parenthetical remarks in Goffman's (1981) descriptive account of the academic lecture. Their objective is "to recommend a mode of procedure or let the audience know what to expect" (Vande Kopple 1985: 85). Contextual comments characterize spoken rather than written discourse, which partly explains why they emerge as particularly interesting to reflect upon in this study.

A manual search of the whole corpus for contextual comments yielded 17 instances. Their realization in TLC is both different from the ones found in the literature on academic lectures in native settings and similar to some of the occurrences in data obtained in contexts similar to the present one. Differences and similarities with those various contexts can be noted when comparing extract (1) with extracts (2) and (3) below.

(1) <l_11> malcolm cowley he said the following words <lecturer dictates>indeed indeed <lecturer dictates> i quote here <lecturer dictates>indeed comma <lecturer dictates>these are one of the **rare** moments where I dictate in fact <lecturer smiles> <students laugh> I am **never** happy I feel **frustrated** all the time but er it is the strategy that we'll we'll opt for and it's for your er</l> <ss> <benefit></s>

<l_11>for your benefit yes er in order to you know try to make you yourselves help in the process of er building this so called course in fact er alright <lecturer dictates></l> <Lit-11-02-A>

(2) I'll move this slide a little bit so you can see better… (Luukka 1994: 80)
You might wish to read the last section first (Crismore *et al.* 1993: 46)

(3) <what I'm speaking is almost English more or less if you neglect the accent the rest should be more or less standard English> (Molino 2018: 946)

Extract (1) illustrates an instance of a contextual comment in TLC where the lecturer justifies his/her use of a dictation strategy. In this extract (as well as in all instances of contextual comments in TLC), contextual comments take the form of a relatively extended talk where the lecturer shares his/her own observations and evaluation of a particular event occurring during the lecture. The lecturer wants to prevent the students from conceiving the course as a mere imparting of information. Indeed, he/she implicitly advises the students against copying verbatim what he/she is dictating during exams. Dictation is negatively viewed in the Tunisian academic context because it is reminiscent of the traditional role of the teacher/lecturer as pouring information to the students' minds, of a view of lectures as information transfer only, and of a perception of the lecturer as the one who holds knowledge. The adjectives *rare* and *frustrated* as well as the adverb *never* are employed by the lecturer to distance him/herself from this technique. The contextual comment also reflects the novice-expert relationship that typically characterizes the academic lecture genre. In the extracts under (2), the commentaries are relatively short and concern the practical management of the materials used during the lecture. They are in line with Vande Kopple's definition of commentaries cited above. In extract (3), however, a clear resemblance with extract (1) can be noted with respect to the nature and motivation of the contextual comment used. Indeed, the lecturer in this extract is referring to the quality of his/her English as somehow departing from the standard form. Molino (2018) argues that the comment acts as a kind of self-protective strategy and the same argument may also apply for the extract from TLC as the lecturer does not want to be negatively perceived as merely pouring information into students as empty vessels.

A second aspect of evaluation is assessing students' knowledge. In <Ling-07-02-B>, 19 steps were coded as evaluating students' knowledge as compared to 51 in <Civ-09-02-A>. Conversely, there is a low proportion of lecture evaluation by the two lecturers, specifically one for <Civ-09-02-A> and seven for <Ling-07-02-B>. Evaluation of lecture points is reflected in sub-functions such as show importance of content and indicating scope. Extracts (4) and (5) illustrate the functions of evaluating students' knowledge and lecture evaluation, respectively.

(4) <l_07> okay classes are of four hours a day</l>
<presenter> <and six days a week></presenter>
<l_07> and six days a week [.] does this remind you of something in Tunisia?
<ss> <no></ss>
<l_07> primary school [.] primary school especially at the [.] o think <unintelligible token="1"/> the first and the second and third primary school this is what pupils do [.] it's five to six days four hours a day of teaching here okay? [.]</l> <Ling-07-02-B>

(5) <l_09> … so they asked for a bill of rights <lecturer writes on the white board> to be added to the constitution [.] what the federalist did is something else they wrote [.] </lecturer writes on the white board> documents commonly known as the federalist papers [.] those documents were <lecturer writes on the white board> a propaganda of federalism it means they talked about the great advantage advantages of federalism okay? so federalists wrote the federalist papers to support the American constitution the antifederalists asked for demanded it's more than asked they demanded <foreign>exiger</foreign> that a bill of rights to be added to the?</l>
<ss> <responses></ss>

> <l_09> constitution don't forget that okay federalist federalist papers antifederalists a bill of rights [..] clear so far? </l> <Civ-09-02-A>

In extract (4), the background knowledge marker *does this remind you of something in Tunisia?* enquires about the students' background knowledge. In doing so, the purpose is to relate the principles and concepts associated with suggestopedia to the students' own educational context. In extract (5), the use of the importance marker *don't forget that* focuses students' attention on the need to associate two important documents in American history, that is, *the Bill of Rights* and *the federalist papers*, to their respective political parties.

The discrepancy between the rhetorical functions of students' knowledge evaluation and lecture evaluation echoes those of Lee (2009) and Cheng (2012) who found that moves pertaining to lecture evaluation are optional. Similarly, Deroey (2017) found that importance markers are less frequent in authentic lectures than in EAP materials. Lecture evaluation refers to the lecturer's weighing of the points he/she makes in terms of importance or relevance. The highly infrequent units displaying evaluation of lecture points as compared to the pervasiveness of units reflecting evaluation of language may suggest that lecture comprehension is pursued at a more local, lexico-grammatical rather than at a global, discourse level of language. More examples and analysis of evaluation as a metadiscursive function are provided in the following section.

5.3. Context-specific metadiscursive functions

A closer inspection of the metadiscursive functions drawn from the present corpus-driven study led to identifying context-specific metadiscourse strategies that were adopted by the two lecturers. One such strategy is evaluating the students' linguistic knowledge through the use of metalinguistic comments (Ädel 2006). These comments are considered metadiscursive because the lecturer does not expand on the propositional meaning of the utterance. Their use expresses the desire to ensure that the students can process the incoming content. Although Ädel (2006: 109) discusses metalinguistic comments in relation to their use by Swedish students of English as "a filler strategy" and as a strategy that reflects non-native speakers' awareness of the situation of writing, the term 'metalinguistic comment' is adopted in this paper because it bears resemblance to those found in TLC as illustrated in extract (6).

(6) <l_09> yes they meet in a conference committee <lecturer writes on the white board> what's the role of a conference committee? </l>
<s><response></s>
<l_09> to iron out <.> differences <lecturer writes on the white board> <lecturer moves arm from right to left> to iron out to try to find a compromise a middle way solution </l> <Civ-09-02-A>

The lecturer in this extract uses both a gesture and a paraphrase in order to define the verb *to iron out*. In Flowerdew's (1992) taxonomy of definitions, the two strategies are referred to as substitution accompanied by visual support. The realization of the metalinguistic comment in extract (6) demonstrates the role that multimodal corpora play in appreciating the way the lecturer combines verbal and non-verbal strategies in order to assist students in following the lecture. In doing so, the lecturer is aware that vocabulary may hinder the comprehension of content. Hence, he/she strives to anticipate and address this issue before moving on to the delivery of academic content.

Metalinguistic comments have been identified under the labels 'low focus definitions' or 'embedded definitions', "which are not the focal point of the information" (Flowerdew 1992: 209) and are "incidental to the logical structure of the lecture" (Jackson and Bilton 1994: 73).[3] Interestingly, their frequency was reportedly greater in non-native contexts like Oman (Flowerdew 1992; Jackson and Bilton 1994), than in contexts where students' English proficiency is considered very high, like Canada (Lessard-Clouston 2009). In the metadiscourse literature, Vande Kopple's (1985) taxonomy of metadiscourse comprises the category 'glosses' that is similar to metalinguistic comments. However, this category, as realized in TLC, was rarely identified in current analyses of corpora, which may be explained by the relatively few studies conducted on academic lectures in an EFL context, a fact which may have made the marker go unnoticed. Moreover, most metadiscourse research was conducted in native and even Content and Language Integrated Learning contexts where the relationship between lecturers and students is considered to be native-to-native, despite the fact that one or both parties are non-native speakers of English. In those contexts, it is possible that metalinguistic comments as they were identified in TLC were simply not

[3] Please, note that definitions of technical and semi-technical words or phrases are included under the rhetorical function 'define content', which is a discourse rather than a metadiscourse function. For examples, see Appendix 3.

used and hence the analyst could not detect them. The aforementioned findings substantiate the need for a function and context-driven approach to corpus annotation.

A second context-specific strategy found is the use of lecturer evaluation markers. This is an evaluative metadiscursive device where the lecturers assess the students' performance to ensure that they appropriately applied theory and/or concepts introduced at an earlier phase of the lecture to case studies. These markers are metadiscursive, since they reflect the direct expression of stance by the lecturer towards the students' discourse. In extract (7) from <Ling-07-02-B>, the lecture evaluation marker takes the form of extended comments where the lecturer evaluates the way the micro-teaching session was carried out by the student.

> (7) <l_07> … okay so you tried the maximum to apply things that are related to the natural approach at especially the first part in a natural approach of course there are other stages that [.] should be included if we have had more time of course but for the start especially for usually beginners we said this is a good strategy to make children of course there is usually a problem a remark that i usually repeat is that you do not adapt yourself to the level you are taught [.] normally you don't know things but you usually answer correct answers … of course you are not er [.] normally students at that level [.] they are learning okay? so they should usually expect the teacher to er help them find the names of things okay? to make the difference between healthy unhealthy [.] …you should adapt yourself to the level you are taught and try to help to the person who is actually teaching to teach okay? </l> <Ling-07-02-B>

A rather positive evaluation is highlighted at the beginning of the extract where the lecturer expressed satisfaction with the way the student applied the principles of the Natural Approach in class. This is conducted through the utterance *you tried the maximum to apply.* A negative evaluation was directed to the other students who were playing the role of pupils at an early stage of their English language learning. The negative evaluation is displayed in the use of the noun problem and the following series of negative grammatical constructions, *viz. do not adapt*, and *don't know.* As such, the negative connotations associated with these two language elements can also be exploited to realize evaluation.

The emergence of lecturer evaluation markers is closely connected to the context of the present study. In fact, the Licence-Master-Doctorate reform in Tunisia strongly urged a shift in focus from theory to practice and more classroom space for interactivity. Accordingly, more time was allotted to practicums in most courses within the undergraduate English curriculum at university. Evaluation of the students' understanding of theory through the implementation of practicums coupled with the pressure for more interactivity in lectures has thus given rise to evaluative metadiscourse as reflected in the use of such markers.

5.4. Signaling of metadiscursive functions

When turning to the linguistic signaling of rhetorical functions, and particularly those which are metadiscursive, a few observations can be made. The lecturer in <Civ-09-02-A>, for instance, did not signal the transition between the lecture structuring phase and the content phase by any means. In some other instances, the relationship between two units was not explicitly signaled, as in extract (8), below, where the cause-effect relationship between the two propositions is implied rather than expounded.

(8) <l_09>so the articles of confederation failed they needed a document</l><Civ-09-02-A>

An explicit realization of the cause-effect relationship would be *so* ***because*** *the articles of confederation failed they needed a document.* The absence of signaling, as far as this metadiscursive function is concerned, adds to the value of function-driven analyses of discourse, especially when they are conducted with a pedagogical purpose in mind. Absence of signaling means that students may not notice the relationship between the two propositions, a fact which may affect their comprehension of the ongoing discourse.

In addition to absence of signaling, other issues were found. In <Civ-09-02-A>, the lecture topic was announced non-verbally as the lecturer wrote the lecture title on the board. No verbal iteration accompanied the visual marking of the topic. This finding further highlights the role that multimodal corpora play in data analysis and interpretation. In her study of academic lectures in universities in Spain, Martín del Pozo (2017: 26) made a similar observation commenting that non-native lecturers "need **more overt** [emphasis added] signaling of lecture phases and a wider stylistic variety enabling them to do so." This is important, since part of the lecturers' role is to produce language that develops the language competence of the students not only at the lexico-grammatical level, but also at the discourse and pragmatic levels. Students should indeed be exposed to models that would encourage them to structure their language productions and develop their academic literacy.

6. Conclusion

This study reports on the compilation of the *Tunisian Lecture Corpus* and provides tools for the analysis of similar corpora for researchers desiring to explore lectures in their own contexts. One major limitation concerns data collection and the absence of follow-up interviews with the lecturers in this study. This kind of data could have provided useful insights into the participants' own perceptions of their pedagogical and linguistic behaviors. Notwithstanding this limitation, the present research contributed to the provision of theoretical, analytical, and methodological tools that can be used to design and approach a corpus of academic lectures. These are the SFL-ESP integrated framework of genre analysis, a corpus-assisted model of genre analysis, and a coding scheme which can serve as a diagnostic tool to approach a corpus when no research questions or hypotheses have been pre-set.

The study yielded four important findings. The first is the discontinuous and cyclical nature of the rhetorical functions in TLC, which is in line with other findings in the literature. The second is the dominance of the metadiscursive functions in the present corpus, particularly structuring when compared to the discourse functions. A third set of findings relates to the new meanings that some metadiscursive categories acquired when compared to other data. This is the case of contextual comments and metalinguistic comments. Both findings reflect the significant role that context plays in shaping the different meanings and linguistic realizations that rhetorical functions have, as well as the value of researching genres in under-investigated contexts. Finally, there were some issues relating to the linguistic signaling of some metadiscursive functions as reflected in the absence of (verbal) signaling for the topic in one of the lectures investigated (Bouziri forthcoming). Given that metadiscursive functions and their realizations are particularly highlighted in the present corpus-driven study, a corpus-based study has been set up to further investigate the use of metadiscourse in TLC (Bouziri 2019).

Important implications of this study are drawn. Firstly, there is a necessity for designing multimodal corpora of lectures which would account for the use of non-verbal strategies. Indeed, the findings of this study highlighted the way some rhetorical functions were realized non-verbally. Along with verbal strategies, these contribute to fulfill the lecturers' pedagogical goals. Secondly, exploring a corpus using a genre analysis framework means that the corpus needs to be human readable. In this respect, the present study provides a coding scheme that could be used as a diagnostic tool. It may be tempting to embark in corpus alignment or the preparation of XML files as one way to make the corpus ready for automatic extractions of some linguistic forms via corpus tools. However, this approach may not be effective for two reasons. The first is that such files are difficult to read for the analyst who is interested in the macro-level and manual discourse coding of lectures. As it was the case for this study, it is important to first start by exploring a sub-set of the corpus using a basic transcription system (cf. Appendix 1) and a manual analysis before deciding to invest time in such endeavors. A second reason is that the search for frequent words or multiword expressions may not yield useful results which would capture special features of the lectures. Again, this stems from the highly idiosyncratic nature of the lecture genre.

Thirdly, a corpus-assisted genre analysis is relevant when corpus studies are conducted with pedagogical objectives in mind. This type of analysis becomes even more significant when conducted in non-native contexts and on genres exhibiting a high degree of variability as it is the case of the academic lecture genre. Such a function-driven approach holds the potential of unveiling rhetorical functions worthy of further research and occasionally their non-canonical realizations. Furthermore, adopting a top-down approach in this study led to uncovering functions which were not signaled or were context-specific. Pedagogically, this is important when it comes to designing professional development programs for lecturers whose aim is to raise their awareness of the potential difficulties that their students may encounter when attending their lectures. It is also significant for the design of local academic materials that would integrate such strategies in order to better reflect the type of discourse that students in Tunisia are exposed to. Despite the difficulty of their implementation and their use of small corpora, macro-level discourse approaches to the lecture genre are rewarding when pedagogical applications are the ultimate objective.

References

Abdesslem, Habib. 1987. *An Analysis of Foreign Language Lesson Discourse: With special Reference to the Teaching of English in Tunisian Secondary Schools*. Sheffield: The University of Sheffield thesis.

Abdesslem, Habib and Hassan Costello. 2018. Introductions in locally published research articles in linguistics: Towards a syntagmatics of moves. *Arab Journal of Applied Linguistics* 3/1: 5–46.

Ädel, Annelie. 2006. *Metadiscourse in L1 and L2 English.* Amsterdam: John Benjamins.

Alsop, Siân and Hilary Nesi. 2014. The pragmatic annotation of a corpus of academic lectures. *LREC*: 1560–1563.

Ballier, Nicolas and Philippe Martin. 2015. Speech annotation of learner corpora. In Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier eds., 107–134.

Bhatia, Vijay K. 2001. Analyzing genre: Some conceptual issues. In Marting Hewings ed. *Academic Writing in Context: Implications and Application*. Birmingham: University of Birmingham Press, 79–92.

Bhatia, Vijay K. 2002. A generic view of academic discourse. In John Flowerdew ed., 21–39.

Biber, Douglas, Ulla Connor and Thomas A. Upton. 2007. *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. Amsterdam: John Benjamins.

Bouziri, Basma. 2019. *A Corpus-assisted Genre Analysis of the Tunisian Lecture Corpus: Focus on Metadiscourse*. Louvain La Neuve: Université Catholique de Louvain thesis.

Bouziri, Basma (Forthcoming). Topic signaling in the *Tunisian Lecture Corpus*. *ESP Today* 8/2: 2–24.

Callies, Marcus. 2015. Learner corpus methodology. In Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier eds., 35–56.

Camiciottoli, Belinda. 2008. Interaction in academic lectures vs. written text materials: The case of questions. *Journal of Pragmatics* 40/7: 1216–1231.

Cheng, Stephanie W. 2012. "That's it for today:" Academic lecture closings and the impact of class size. *English for Specific Purposes* 31/4: 234–248.

Crismore, Avon, Raija Markkanen and Margaret S. Steffensen. 1993. Metadiscourse in persuasive writing: A study of texts written by American and Finnish university students. *Written Communication* 10/1: 39–71.

Deroey, Katrin L. B. and Miriam Taverniers. 2011. A corpus-based study of lecture functions. *Moderna Språk* 2: 1–22.

Deroey, Katrin L. B. 2017. How representative are EAP listening books of real lectures? In Jenny Kemp ed. *2015 BALEAP Conference Proceedings: EAP in a rapidly Changing Landscape: Issues, Challenges and Solutions*. Reading: Garnet Publishing.

Dudley-Evans, Tony. 1994. Variations in the discourse patterns favored by different disciplines and their pedagogical implications. In John Flowerdew ed., 146–158.

Duszak, Anna. 1994. Academic discourse and intellectual styles. *Journal of Pragmatics* 21/3: 291–313.

Flowerdew, John. 1992. Definitions in science lectures. *Applied Linguistics* 13/2: 202–221.

Flowerdew, John ed. 1994. *Academic Listening: Research Perspective.* Cambridge: Cambridge University Press

Flowerdew, John ed. 2002. *Academic Discouse*. London: Longman.

Flowerdew, John. 2002. Genre in the classroom: A linguistic approach. In Anne Johns ed. *Genre in the Classroom: Multiple Perspectives.* London: Lawrence Erlbaum Associates, 91–120.

Flowerdew, Lynn. 2004. The argument for using English specialized corpora to understand academic and professional language. In Ulla Connor and Thomas A. Upton eds. *Discourse in the Professions: Perspectives from Corpus Linguistics*. Amsterdam: John Benjamins, 11–33.

Flowerdew, John. 2013. *Discourse in English Language Education*. London: Routledge.

Goffman, Erving. 1981. *Forms of Talk*. Philadelphia: University of Pennsylvania Press.

Granger, Sylviane, Gaëtanelle Gilquin and Fanny Meunier eds. 2015. *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

Gregory, Michael. 2002. Phasal analysis within communication linguistics. In Michael Cummings, Peter Fries, David Lockwood and William Spruiell eds. *Relations and Functions within and around Language*. London: Continuum, 316–345.

Hyland, Ken. 2002. Activity and evaluation: Reporting practices in academic writing. In John Flowerdew ed*.*, 115–130.

Hyon, Sunny. 1996. Genre in three traditions: Implications for ESL. *TESOL Quarterly* 30/4: 693–722.

Jackson, Jane and Linda Bilton. 1994. Stylistic variations in science lectures: Teaching vocabulary. *English for Specific Purposes* 13/1: 61–80.

Kwan, Becky S. C. 2006. The schematic structure of literature reviews in doctoral theses of applied linguistics. *English for Specific Purposes* 25/1: 30–55.

Lee, Joseph. 2009. Size matters: An exploratory comparison of small and large-class university lecture introductions. *English for Specific Purposes* 28/1: 42–57.

Lessard-Clouston, Michael. 2009. Definitions in theology lectures: Implications for vocabulary learning. *The Asian ESP Journal* 5/1: 7–22.

Luukka, Miina. Riitta. 1994. Metadiscourse in academic texts. In Britt-Louise Gunnarson, Per Linell and Bengt Nordberg eds. *Text and Talk in Professional Contexts*. Uppsala: The Swedish Association of Applied Linguistics.

Martín del Pozo, María Ángeles. 2017. Training teachers for English Medium Instruction: Lessons from research on second language listening comprehension. *Revista de Lingüística y Lenguas Aplicadas* 12/1: 55–63.

Molino, Alessandra. 2018. 'What I'm speaking is almost English…': A corpus-based study of metadiscourse in English Medium Lectures at an Italian university. *Educational Sciences: Theory and Practice* 18/4: 935–956.

Nesi, Hilary. 2008. Corpora and EAP. In *LSP: Interfacing Language with other Realms: Proceedings of the 6th Languages for Specific Purposes International Seminar*. Johor Bahru: Universiti Teknologi Malaysia. https://warwick.ac.uk/fac/soc/al/research/collections/bawe/papers/corpora_and_eap.pdf

O'Donnell, Mike. 2017. *UAM CorpusTool*. Madrid: Universidad Autónoma de Madrid.

Palmer-Silveira, Juan Carlos. 2004. Delivery strategies in classroom lectures: Organising the message. In Pilar Garcés, Reyes Gómez, Lucía Fernández and Manuel Padilla eds. *Current Trends in Intercultural, Cognitive and Social Pragmatics*. Sevilla: Editorial Kronos, 97–115.

Partington, Alan. 2004. Corpus-assisted discourse studies. In Alan Partington, John Morley and Louann Haarman eds. *Corpora and Discourse*. Berlin: Peter Lang, 9–18.

Swales, John. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Swales, John. 2004. *Research Genres: Explorations and Applications*. Cambridge: Cambridge University Press.

Thompson, Susan. 1994. Frameworks and contexts: A genre-based approach to analyzing lecture introductions. *English for Specific Purposes* 13/2: 171–186.

Thompson, Susan. 2003. Text-structuring metadiscourse, intonation and the signaling of organization in academic lectures. *Journal of English for Academic Purposes* 2/1: 5–20.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Touati, Walid. 2004. *Native vs. Non-native Treatment of Oral Errors in an EFL Context.* Tunisia*:* Institut Supérieur des Langues de Tunis dissertation.

Vande Kopple, William J. 1985. Some exploratory discourse on metadiscourse. *College Composition and Communication* 36/1: 82–93.

Wu, Shuxuan. 2013. Discourse structure and listening comprehension of English academic lectures. *Theory and Practice in Language Studies* 3/9: 1705–1709.

Yaakob, Salmah. 2013. *A Genre Analysis and Corpus-based Study of University Lecture Introductions*. Birmingham: The University of Birmingham dissertation.

Yin, Bin. 2016. An exploratory genre analysis of three graduate degree research proposals in applied linguistics. *Functional Linguistics* 3/7: 1–28

Young, Lynn. 1990. *Language as Behavior, Language as Code: A Study of Academic English*. Amsterdam: John Benjamins.

Young, Lynn. 1994. University lectures-macro-structure and micro-features. In John Flowerdew ed., 169–178.

Corresponding author
Basma Bouziri
University of Gabés
Institute of Arts and Crafts of Tataouine
Department of Education
Cité Mahrajène
3200 Tataouine
Tunisia
Email: basma.bouziri@isamt.rnu.tn

received: May 2020
accepted: September 2020

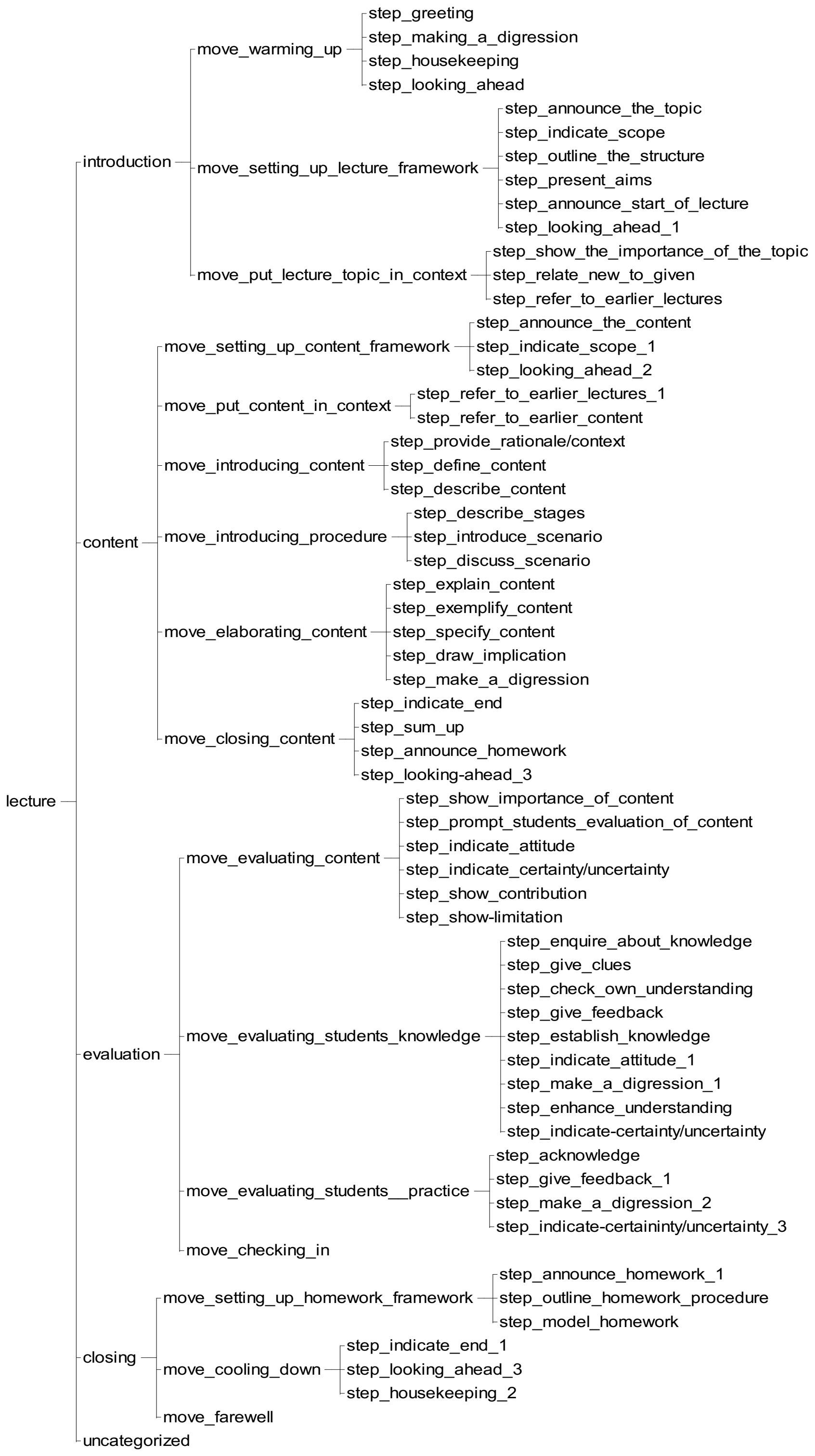
APPENDICES

Appendix 1: Transcription guidelines[4]

ELEMENT	TRANSCRIPTION CONVENTION
Backchannels	Examples include: *oh, oops, really, okay, yeah, uhm, no, yes, right*
Contextual events	-Teacher does not speak while he writes <the lecturer writes on the white board dur= 5 secs> -Teacher speaks while he writes <the lecturer writes on the white board>text</the lecturer writes on the white board>
Error correction	<error corr=this>these</error corr=this>
Fillers	Fillers like er, err, erm, mm are transcribed
Participants	Lecturer is transcribed as 'l' followed by the code attributed to him/her. Example: <l_01> <s> refers to one student. <ss> refers to a group of students. <presenter> refers to a student making a presentation. <presenters> refer to two or more students making a presentation.
Pauses	[.] [..] [pause dur=7 secs]
Sounds	<sound= 'tch'/>
Translation	Translation is included between a double parenthesis <foreign>Oay</foreign> ((yes))
Unintelligible tokens	<unintelligible token='1'/>

[4] A more detailed version of the transcription guidelines is available at https://www.iris-database.org/iris/app/home/detail?id=york:938327

Appendix 2: Coding scheme for rhetorical functions



- lecture
  - introduction
    - move_warming_up
      - step_greeting
      - step_making_a_digression
      - step_housekeeping
      - step_looking_ahead
    - move_setting_up_lecture_framework
      - step_announce_the_topic
      - step_indicate_scope
      - step_outline_the_structure
      - step_present_aims
      - step_announce_start_of_lecture
      - step_looking_ahead_1
    - move_put_lecture_topic_in_context
      - step_show_the_importance_of_the_topic
      - step_relate_new_to_given
      - step_refer_to_earlier_lectures
  - content
    - move_setting_up_content_framework
      - step_announce_the_content
      - step_indicate_scope_1
      - step_looking_ahead_2
    - move_put_content_in_context
      - step_refer_to_earlier_lectures_1
      - step_refer_to_earlier_content
    - move_introducing_content
      - step_provide_rationale/context
      - step_define_content
      - step_describe_content
    - move_introducing_procedure
      - step_describe_stages
      - step_introduce_scenario
      - step_discuss_scenario
    - move_elaborating_content
      - step_explain_content
      - step_exemplify_content
      - step_specify_content
      - step_draw_implication
      - step_make_a_digression
    - move_closing_content
      - step_indicate_end
      - step_sum_up
      - step_announce_homework
      - step_looking-ahead_3
  - evaluation
    - move_evaluating_content
      - step_show_importance_of_content
      - step_prompt_students_evaluation_of_content
      - step_indicate_attitude
      - step_indicate_certainty/uncertainty
      - step_show_contribution
      - step_show-limitation
    - move_evaluating_students_knowledge
      - step_enquire_about_knowledge
      - step_give_clues
      - step_check_own_understanding
      - step_give_feedback
      - step_establish_knowledge
      - step_indicate_attitude_1
      - step_make_a_digression_1
      - step_enhance_understanding
      - step_indicate-certainty/uncertainty
    - move_evaluating_students__practice
      - step_acknowledge
      - step_give_feedback_1
      - step_make_a_digression_2
      - step_indicate-certaininty/uncertainty_3
    - move_checking_in
  - closing
    - move_setting_up_homework_framework
      - step_announce_homework_1
      - step_outline_homework_procedure
      - step_model_homework
    - move_cooling_down
      - step_indicate_end_1
      - step_looking_ahead_3
      - step_housekeeping_2
    - move_farewell
  - uncategorized

Appendix 3: Examples of rhetorical functions

RETHORICAL FUNCTION	EXAMPLES
STRUCTURING	
FRAMING	
SETTING_UP_LECTURE_FRAMEWORK	
Announce_the_topic	<l_7>today we're going to deal with a method which is completely different from the others [.]</l>
Outline_the_structure	None
Present_aims	None
Announce_start_of_lecture	None
Looking_ahead	None
SETTING_UP_CONTENT_FRAMEWORK	
Announce_content	<l_07>so we start first with the er [.] background yes</l>
Looking_ahead	<l_07>a lot of arts and drama [..] in the teaching **and we will see later on how can this happen during the process**<l>
REVIEW _CONTENT	
Indicate_end	<l_07> **so this is the first thing** the second this is</l>
Sum_up	<l_07>so this is the first thng you write [.] the first observation of lozanov is [.] that learners do not use more than ten percent of their brain capacity under traditional methods of teaching [.] </l>
REVIEW_ LECTURE	
Indicate_end	<l_09>okay so let's stop here</l>
Housekeeping	<l_09>so did I tell you about the test? <s> <responses> <l_09>it will be the first week after the holidays which means the thirtieth of march</l>
Looking ahead	<l_07>I see you next time [.] for multiple intelligence lessons</l>
CONTEXTUALIZING	
PUT_LECTURE TOPIC_IN_CONTEXT	
Refer_to_earlier_lectures	<l_09>so we already focused in the first sessions on different periods of american history we talked about the colonial period [.]
PUT_CONTENT_IN_CONTEXT	
Refer_to_earlier_content	<l_07>remember that pupils said that the system is not fair because you teach us in the same way even if we have different? levels</l>
Provide_rationale/context	<l_07>teachers do not give homework because he discovered also through a questionnaire and the studies that he did that one of the problems with school is? Homework</l>
ELABORATING	
Explain_content	<l_07>it's a method of teaching that has an objective to? [.] lower or eliminate all the psychological barriers that pupils develop during their learning process **and hence what can we do? We can maximize what now? their use of their brain capacities [.] if we lower these psychological barriers students are going to use more their cognitive skills**</l>
Exemplify_content	<l_07>there are classrooms which are special for teaching english okay? and teachers make an effort to decorate them..with different things that are related to the e,glish culture maps I don't know the maps of Britain of the US some poems famous famous famous sayings different things pictures in which you find the parts of the body the different flags</l>
Specify_content	<l_07>this is his diagnosis of the state of learners while they are taught according to the conventional teaching methods **of course here we talk about grammar translation method audiolingual err method of teaching the cll the different types of methods that existed before when pupils are taught using the methods**</l>
Draw_implication	<l_07>remember that we said that he discovered that this brain capacity is limited to? [.] ten percent **so we try to? maximize that through lowering these psychological barriers** okay?</l>

RETHORICAL FUNCTION	EXAMPLES
DEFINING AND DESCRIBING	
Define_content	<l_07> so can you tell us what is desuggestion?</l> <s><student resumes presentation dur=12 secs></s> <l-07>okay meaning offering options and proposals</l>
Describe_content	<l_07> a suggestopedic class should be like this the first thing as you said comfortable [.] what? <s>comments</s> <l_07> okay so comfortable class so we should have a comfortable environment yes?so this is important the chairs are arranged in a semi-circle </l>
EVALUATING_ LECTURE_ CONTENT	
Show_importance_of_content	<l_07>and it is very important [.] to highlight that this person is a psychotherapist because all of this is going to influence</l>
Indicate_Scope	<l_09>okay so let's move to the most important part which is the constitution itself</l>
EVALUATING_KNOWLEDGE	
Enquire_about_ knowledge	<_07> georgi lozanov [.] who is actually what? <s>response</s> <l_07> a psychotherapist basically and educator [.] what is a psychotherapist?</l>
Give_clues	<l_09>what is the meaning of a census? **sta? [.]** census? statistics which are made every ten years…<:l>
Establish_knowledge	<l_07>lowering means of course diminishing the psychological barriers</l>
Indicate_attitude	<l_09>no i personally i <foreign>belEaks</foreign> ((on the contrary)) he is the least charismatic of all the russian presidents er</l>
Indicate certainty/uncertainty	<l_07>**I don't know whether you have ever gone to a psychotherapist** but in a cabinet of a psychotherapist there are two key features</l>
Give_feedback	<l_07>so you tried the maximum to apply things that are related to the natural approach of course there are other stages that [.] should be included</l>
Enhance_understanding	<l_07>mozart bethoven vangelis bach all of these are type of baroque music<:l>

RiCL Research in Corpus Linguistics

Masked by annotation: Minor declarative complementizers in parsed corpora of historical English

María José López-Couso – Belén Méndez-Naya
University of Santiago de Compostela / Spain

Abstract – This article discusses some of the potential problems derived from the syntactic annotation of historical corpora, especially in connection with low-frequency phenomena. By way of illustration, we examine the parsing scheme used in the *Penn Parsed Corpora of Historical English* (PPCHE) for clauses introduced by so-called 'minor declarative complementizers', originally adverbial links which come to be occasionally used in complementizer function. We show that the functional similarities between canonical declarative complement clauses introduced by the major declarative links *that* and zero and those headed by minor declarative complementizers are not captured by the PPCHE parsing, where the latter constructions are not tagged as complement clauses, but rather as adverbial clauses. The examples discussed reveal that, despite the obvious advantages of parsed corpora, annotation may sometimes mask interesting linguistic facts.

Keywords – annotation; parsing; English historical corpora; minor declarative complementizers; indeterminacy

1. INTRODUCTION[1]

Since the advent of the first computerized corpora in the 1960s (e.g. the compilation of the pioneering *Brown Corpus* by W. Nelson Francis and Henry Kučera), corpus linguistics has experienced exponential growth. In just half a century we have witnessed the creation of an impressive range of corpora, written, spoken and multimodal, small collections of data alongside mega-corpora and reference corpora side by side with different types of specialized corpora.

One of the milestones of modern corpus linguistics has undoubtedly been the development of various types of annotation systems, such as tagging and parsing, which

[1] The research reported on in this paper has been funded by the Spanish Ministry of Economy, Industry and Competitiveness (National Programme for Excellence in Scientific and Technical Research), Grant FFI2017-86884-P; and the Regional Government of Galicia, Consellería de Cultura, Educación e Universidade, Grant ED431B 2020/01. Thanks are also due to two anonymous reviewers for their helpful comments.

Research in Corpus Linguistics 8/2: 133–158 (2020). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.08.02.07

add significantly to the potentials of earlier collections of raw data, allowing the analyst to run quicker and more effective searches. It must be acknowledged, however, that linguistic annotation also has certain disadvantages, mostly because any annotation system implies the acceptance of particular theoretical premises, no matter how inclusive the annotators claim to be. Even the identification and tagging of a basic grammatical category such as 'preposition' can be controversial (see Huddleston and Pullum *et al*. 2002: 598–601, who include traditional subordinating conjunctions, such as *since* or *because*, under the category 'preposition', *contra* Quirk *et al*. 1985: 658–661). Linguistic annotation of historical material may be even more problematic, since parsing has to account for language change, especially if the annotation system is supposed to hold for successive stages in the history of a given language. Consider, for instance, the intrinsic difficulties in the annotation of so-called 'bridging contexts' in grammaticalization (Heine 2002), where parsing may mask cases of potential syntactic and semantic indeterminacy or ambiguity, which are central to our understanding of linguistic change.

Further problems with annotation may arise in the treatment of low-frequency phenomena, which tend to be overlooked. One of these low-frequency features is so-called 'minor declarative complementizers', a category which we have analyzed in depth for English from both a theoretical and a diachronic point of view (for an overview, see López-Couso and Méndez-Naya 2015). Minor declarative complementizers are connectives whose main function is that of marking various kinds of adverbial relations (e.g. condition, concession, purpose, comparison, etc.), but which also serve a secondary function to introduce finite complement clauses, as equivalents (or near equivalents) of the major declarative complementizers *that* and zero. Examples of such minor complementizers in English include *if*, *though*, *as if*, *as though*, *like*, *lest* and *but* (see examples (3a)–(3e) in Section 3 below).

In this article we draw attention to some of these minor declarative complementizers by examining the way in which they are annotated in the *Penn Parsed Corpora of Historical English* (Kroch and Taylor 2000; Kroch *et al*. 2004, 2016), with the aim of checking whether parsing overlooks diachronic facts and/or masks diachronic developments. The outline of the discussion is as follows. In Section 2 we introduce the *Penn Parsed Corpora of Historical English* and their common annotation scheme. Section 3, in turn, summarizes the most relevant information about minor declarative complementizers from our earlier research, focusing on the various structural and semantic criteria which lead us to consider the clauses introduced by these connectives as complements rather than as adjuncts. Then, in Section 4 we show how such clauses are annotated in the corpora, as compared to canonical cases of finite complement clauses introduced by the major declarative complement-clause links *that* and zero. Finally, Section 5 offers some concluding remarks.

2. THE *PENN PARSED CORPORA OF HISTORICAL ENGLISH* AND THEIR ANNOTATION SYSTEM

A major landmark in the history of English historical corpora was the release in 1991 of the *Helsinki Corpus of English Texts* (HC), a project launched by Matti Rissanen and his collaborators at the Department of English of the University of Helsinki back in 1984. The HC is a 1.5 million-word corpus which contains text material from the time of the earliest written records of English in the eighth century up to the first decade of the eighteenth century, representing a wide range of genres, both formal (e.g. philosophical treatises) and informal (e.g. comedies).[2] Almost three decades after its publication and despite its small size according to modern standards, the HC still remains an excellent resource for corpus-based research on the long diachrony of English and is still successfully used world-wide as a “diagnostic” corpus (Rissanen 2008: 59) for the analysis of processes of language change taking place in the Old, Middle and Early Modern English periods.

In order to expand the potentialities of the original ‘raw’ version of the HC, a number of complementary corpora have appeared over the last couple of decades or so. Of special relevance for our purposes in this article are the annotated (tagged or parsed) editions for various historical sub-periods developed by a team of scholars based at the University of Pennsylvania and at the University of York, a project aimed at producing syntactically annotated corpora for the different stages in the history of the English language. Though based on the raw version of the HC, these corpora contain considerably larger text samples than those in the HC, together with some new material not available in the original corpus. The advantages of these complementary annotated corpora over their raw counterparts are more than evident: in addition to searches for simple words or word strings, they allow searching for syntactic constructions, including empty or covert

[2] For full details about the HC, see the third edition of the manual by Kytö (1996) and the corresponding entry for the corpus in the *Corpus Resource Database* (CoRD) at http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/.

categories, such as empty subjects or zero complementizers, thus conveniently facilitating the analyst's task, especially when high-frequency phenomena are in focus.

At present, the *Penn Parsed Corpora of Historical English* (PPCHE) comprise the following datasets:

- *York-Helsinki Parsed Corpus of Old English Poetry* (YCOEP)
- *York-Toronto-Helsinki Parsed Corpus of Old English Prose* (YCOE)
- *Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English*
- *Penn-Helsinki Parsed Corpus of Middle English, second edition* (PPCME2)
- *Penn-Helsinki Parsed Corpus of Early Modern English* (PPCEME)
- *York-Helsinki Parsed Corpus of Early English Correspondence* (PCEEC)
- *Penn Parsed Corpus of Modern British English, second edition* (PPCMBE2)

The PPCHE are presented in three different formats: simple or raw text, part-of-speech (POS) tagged and parsed text, which combines both POS and syntactic annotation (treebanks). For the annotated versions of the PPCHE, the compilers adopted a simplified version of the Principles and Parameters theory. The annotation scheme used is ultimately based on the system developed for the *Penn Treebank*, a corpus of over 4.5 million words of American English (Marcus *et al*. 1993). This system was adapted to historical material by Ann Taylor and Anthony Kroch for the second edition of the *Penn-Helsinki Parsed Corpus of Middle English* (Kroch and Taylor 2000) and was then revised for the 2016 update of the PPCHE.[3]

In the annotation manual (Santorini 2016) it is made clear that the primary goal of the Penn-Helsinki annotation system is to facilitate automated searches, "not to give the correct linguistic analysis of each sentence." In other words, practical purposes are clearly privileged over grammatical ones. In addition, Santorini mentions that in the annotation process subjective judgements have been avoided "since they are extremely error-prone." This explains why distinctions such as, for instance, adjectival vs. verbal passive participles are disregarded. Practical issues also prevail, for instance, in the different

[3] The annotation scheme used for the PPCHE has also been applied to various other historical datasets, both for English (e.g. the *Parsed Corpus of Middle English Poetry*; https://pcmep.net/index.php) and for other languages, including Portuguese, French, Icelandic and Japanese, among others. For a list of corpora sharing the same annotation scheme, see https://www.ling.upenn.edu/ppche/ppche-release-2016/other-corpora.html.

parsing given for the clauses introduced by *because* and variants, depending on whether the form is written together or apart. In the first case the clause following *because* is analyzed as an adverbial clause (CP-ADV), whereas in the second case the clause is treated as a complement clause (CP-THT):

> BECAUSE is treated as a fused form. When it is written together, the clause following it is treated as the CP-ADV complement of the compound head P+N. When it is written apart, the clause following it is treated as a THAT complement of the noun CAUSE.

```
( (IP-MAT (CONJ but)
          (NP-SBJ (NPR$ Balynes) (N oste))
          (MD myght)
          (NEG $not)
          (BE be)
          (VAN lette)
          (RP in)
          (PP (P+N because)
              (CP-ADV (C 0)
                      (IP-SUB (NP-SBJ (PRO he))
                              (HVD had)
                              (NP-OB1 (Q no) (N lady))))))
          (. .))
          (ID CMMALORY,63.2096))


( (IP-MAT (CONJ but)
          (PP (P by)
              (NP (N cause)                        ← by cause that
                  (CP-THT (C 0)
                          (IP-SUB (NP-SBJ (PRO he))
                                  (VBD knewe)
                                  (NEG not)
                                  (NP-OB1 (PRO$ his) (N sheld))))))
          (NP-SBJ (PRO he))
          (VBD demed)
          (CP-THT (C 0)
                  (IP-SUB (NP-SBJ (PRO it))
                          (BED was)
                          (NEG not)
                          (NP-OB1 (PRO he))))
          (. .))
          (ID CMMALORY,68.2300))
```

This parsing, which clearly favors automation, relies exclusively on the expressions' surface structure: *because* is analyzed as a one-word item that governs an adverbial CP, while in the variant *by cause*, the first word is marked as governing the noun phrase headed by the noun *cause*.

As mentioned in Section 1, in contrast to the annotation of contemporary data, the tagging and parsing of historical material pose special challenges to both annotators and corpus users, especially when it comes to the interpretation of items and constructions undergoing processes of language change such as grammaticalization and lexicalization, which imply alterations in the status of a given item or construction over time (e.g. from lexical item to grammatical item; from syntactic construction to lexical item). The parsing system of the PPCHE conveniently tries to accommodate such changing diachronic facts. This applies in particular to differences between the Middle English corpus (PPCME2) and the later corpora. Thus, for example, the annotation reflects the emergence of conjunctions and adverbs out of phrases and clauses, providing a different parsing for the source constructions and for the grammaticalized elements. By way of illustration, consider the convincing explanation provided in the manual for the development of the subordinator *albeit*:

> ALL BE IT (THAT), ALBEIT
>
> In the PPCME2, ALL BE IT (THAT) clauses, like SO BE IT (THAT) clauses, are treated similarly to V1 conditionals. ALL is POS-tagged Q, surrounded by ADVP brackets, and treated as a daughter of CP-ADV. This is not intended as the correct analysis of the construction, but rather to fit in with the annotation of V1 conditionals.

```
( (IP-MAT (CONJ and)
          (PP (P atte)
              (NP (N risyng)
                  (PP (P of)
                      (NP (D the) (N sonne)))))
          (NP-SBJ (PRO I))
          (VBD fond)
          (NP-OB1 (D the) (ADJ secunde) (N degre)
                  (PP (P of)
                      (NP (NPR Aries))))
          (IP-PPL (VAG sittyng)
                  (PP (P upon)
                      (NP (PRO$ myn) (N est) (N orisonte))))
          (, ,)
          (CP-ADV (ADVP (Q all))                          ← ALL BE IT
                  (IP-SUB (BEP be)
                          (NP-SBJ-1 (PRO it))
                          (CP-THT-1 (C that)
                                    (IP-SUB (NP-SBJ (PRO it))
                                            (BEP was)
                                            (ADJP (FP but) (ADJ litel))))))
          (. .))
  (ID CMASTRO,673.C1.364))
```

In the later corpora, ALBEIT (like HOWBEIT) is treated as a unitary adverb (when used absolutely) or as a unitary preposition (when introducing a subordinate clause).

```
(NODE (CP-CAR (WNP-1 (WPRO Which))
            (C 0)
            (IP-SUB (PP (P in)
                        (NP (NP-POS (D the) (N$ kinges))
                             (NS daies)))
                    (, ,)
                    (PP-LFD (P albeit)          ← ALBEIT
                             (CP-ADV (C 0)
                                      (IP-SUB (NP-SBJ (PRO he))
                                              (BED was)
                                              (ADVP (ADV sore))
                                              (VAN ennamored)
                                              (PP (P vpon)
                                                   (NP (PRO her))))))
                    (, ,)
                    (ADVP-RSP (ADV yet))
                    (NP-SBJ-RSP=1 (PRO he))
                    (VBD forbare)
                    (NP-OB1 (PRO her))
      (ID MORERIC,55.118))
```

Table 1 extracted from the manual[4] summarizes the differences in the treatment of *albeit* in the PPCME2 and in the later corpora.

Item	**PPCME2**	**Later corpora**

ALL BE IT,
ALBEIT
(see Concessive clauses)
	Always phrasal.
	`(Q all) (BEP be) (PRO it)`
		Unitary adverb or preposition.
		`(ADV albeit)`
		`(ADV (ADV31 al) (ADV32 be) (ADV33 it))`
		`(P albeit)`
		`(P (P31 al) (P32 be) (P33 it))`

Table 1: The annotation of *albeit* in the PPCME2 and in the later PPCHE corpora

Moreover, the annotators of the PPCHE also acknowledge the existence of ambiguity by explaining alternative analyses, even though the annotation finally opts for a default interpretation, as shown below in the case of the verb *do*:

> In Middle English, DO can be ambiguous between a causative (ECM) main verb and a periphrastic auxiliary. The default in the PPCME2 is to treat ambiguous cases as causative except when a causative reading is impossible. Causative DO dies out in the course of Middle English, and so instances of DO in the later corpora that could in principle be treated as ambiguous and hence causative by default are instead uniformly treated as periphrastic.

Though recognizing ambiguity, practical purposes finally prevail in the annotation used in the PPCHE system. In addition to facilitating the retrieval of examples, this solution has the obvious advantage of enabling the automation of the annotation process.

[4] https://www.ling.upenn.edu/hist-corpora/annotation/index.html

In the remainder of this article we examine another controversial area of syntactic interpretation, namely so-called ‘minor declarative complementizers’, and discuss the way(s) in which such subordinators and the clauses introduced by them are treated in the PPCHE annotation scheme.

3. INTRODUCING MINOR DECLARATIVE COMPLEMENTIZERS

Most Present-day English reference grammars (see Quirk *et al*. 1985: 1047ff; Biber *et al*. 1999: 192ff) distinguish three main classes of subordinate clauses on the basis of their potential functions in the complex sentence: complement clauses, which realize functions that approximate to those of noun phrases; relative clauses, which resemble adjectives in function; and adverbial clauses, which are found in functions more closely associated with adverbial and prepositional phrases, expressing satellite relations and acting as adjuncts or modifiers. These different functional categories of subordinate clauses are introduced by various kinds of markers indicating the type of relating function which exists between the subordinate clause and its corresponding superordinate clause. The two types of subordinate-clause links which are relevant for the present discussion are those introducing complement clauses (i.e. complementizers) and subordinators which mark adverbial clauses of various kinds.

Finite declarative complement clauses are typically introduced by the complementizers *that* or zero, as in (1a)–(1b).

(1a) I noticed ***that*** *he spoke English with an Australian accent*. (from Quirk *et al*. 1985: 1049)

(1b) I know **Ø** *it’s late*. (from Quirk *et al*. 1985: 1049)

In turn, adverbial clauses are normally marked by the presence of different subordinators in clause-initial position. These markers signal the various kinds of semantic relations which may hold between the main clause and the sub-clause, among them time (2a), reason (2b), condition (2c), concession (2d), exception (2e), (negative) purpose (2f), comparison (2g), etc. (see Quirk *et al*. 1985: 1077ff; Kortmann 1997: 79ff; Biber *et al*. 1999: 818ff).

(2a) ***Since*** *I saw her last*, she has dyed her hair. (from Quirk *et al*. 1985: 1078)

(2b) The flowers are growing so well ***because*** *I sprayed them*. (from Quirk *et al*. 1985: 1103)

(2c) ***If*** *Mary had visited her parents yesterday*, she would have known about their problems. (from Kortmann 1997: 85)

(2d) He can walk faster than I can, ***though*** *he is well over eighty*. (adapted from Quirk *et al*. 1985: 1097)

(2e) Rumsfeld: "There is no question **but that** *the invasion would be welcomed*." (COCA, 2006, MAG)

(2f) I sent the children to bed ***lest*** *they (should) hear their parents quarrel*. (from Kortmann 1997: 86)

(2g) He treats me ***as if*** *I am a stranger*. (from Quirk *et al*. 1985: 1110)

Interestingly, some of the adverbial connectives illustrated in (2a)–(2g) may also show a secondary or subsidiary function beyond the domain of adverbial subordination and are also used (or have been used at different stages in the history of English) to introduce finite declarative complements, as equivalents or near-equivalents of the major declarative complementizers *that* and zero. Illustrative examples are given in (3).

(3a) It would be a real comfort to me ***if*** *you would make me feel we belonged to each other*. (ARCHER, 1893pine.d6b; from López-Couso and Méndez-Naya 2014: 98)

(3b) and therfore, ***though*** *I be wrooth and inpaciente*, it is no merveille (c. 1390, Chaucer, Tale of Melibee 232.C2; from López-Couso and Méndez-Naya 2001: 99)

(3c) I don't doubt ***but that*** *she meant it*. (from Huddleston and Pullum *et al*. 2002: 971)

(3d) He suggested offering half to Sir Edward, fearing ***lest*** *"he shall thinke it to good for us and procure it for himselfe, as he served us the last time"*. (Brown G64 S25)

(3e) It seems ***as if*** (***as though***) *we're in a bad situation*, no matter your point of view. (adapted from COCA, 2018, SPOK)

However, while the variation between *that* and zero has been widely examined in the extensive literature on clausal complementation both from a synchronic and from a diachronic perspective (see, among many others, Elsness 1982, 1984; Warner 1982; Fanego 1990; Rissanen 1991; Finegan and Biber 1995; López-Couso 1996; Tagliamonte and Smith 2005; Kaltenböck 2006; Kearns 2007; Torres Cacoullos and Walker 2008), the complementizer function of these subordinators has been largely overlooked.[5] More importantly for the purposes of the present article, this neglect has had serious implications for the way in which these connectives and the clauses introduced by them have been annotated in parsed corpora.

In various publications we have drawn attention to the complementizer use of these originally adverbial links, which we have labelled 'minor declarative complementizers', and have dealt with their origin, development and present-day use. In particular, we have examined *but (that)* (López-Couso and Méndez-Naya 1998), *if* and *though* (López-Couso and Méndez-Naya 2001, 2014), *lest* (López-Couso 2007) and *as if*, *as though* and *like* (López-Couso and Méndez-Naya 2012a; 2012b). In these articles we have argued that even though clauses such as those italicized in (3a)–(3e) above resemble adverbial clauses at first sight, they nevertheless meet a number of criteria which favor a complement analysis. What follows summarizes the discussion of the criteria for complementhood that we proposed in López-Couso and Méndez-Naya (2015).

(a) Licensing. The most central criterion is licensing, inasmuch as complements depend on the presence of a predicate "that licenses them" (Huddleston and Pullum *et al*. 2002: 219). In contrast to the adverbial clauses in (2), the occurrence of the subordinate clauses in (3) requires the presence of a particular kind of predicate. In example (3d), for instance, the clause introduced by *lest* is perfectly grammatical with a predicate of fearing such as the verb *fear*, but it would be ungrammatical with the utterance predicate *say* (see (4)).[6]

(3d) He suggested offering half to Sir Edward, fearing ***lest*** *"he shall thinke it to good for us and procure it for himselfe, as he served us the last time"*.

[5] Minor declarative complementizers are discussed in passing in Lakoff (1968: 69, note 7); Huddleston (1971: 177–178); Warner (1982: 180–185, 221–224); Mitchell (1985: §§1960–1961); Noonan (1985: 104); Quirk *et al*. (1985: 1175, note a); McCawley (1988: 143); Dirven (1989: 134); Fanego (1990: 19–20); Huddleston and Pullum *et al*. (2002: 962, 1151–1152); Dancygier and Sweetser (2005: 229–230) and Taylor and Pang (2008: 130). The comparative complementizers *as if* and *as though* are discussed in greater detail in Bender and Flickinger (1999), Rooryck (2000) and Brook (2014, 2018).

[6] We follow here the classification of semantic predicates proposed by Noonan (1985).

(4) *He suggested offering half to Sir Edward, saying ***lest*** *"he shall thinke it to good for us and procure it for himselfe, as he served us the last time"*.

(b) Obligatoriness. As a direct consequence of licensing, complements are obligatory constituents in clause structure; in other words, their omission would compromise the grammaticality of the sequence. As can be seen, the italicized clauses in (3) are not omissible (see (5)), which clearly supports a complementation analysis for them. By contrast, adverbial clauses, like the concessive *though*-clause in (2d), can easily be left out, as shown in (6). As complements, the sub-clauses in (3) show therefore a higher degree of integration into the corresponding matrices than the adverbial clauses in (2).

(3e) It seems ***as if*** *we're in a bad situation*.

(5) *It seems.

(2d) He can walk faster than I can, ***though*** *he is well over eighty*.

(6) He can walk faster than I can.

(c) Replacement by unambiguous declarative complement clauses. Further evidence in favor of the complement status of the subordinate clauses in (3) is provided by their ability to be replaced by prototypical declarative complement clauses, either finite or non-finite, "without any perceptible change of meaning" (Huddleston and Pullum *et al.* 2002: 962). Consider, for instance, the alternatives provided in (7) and (8) for the sequences in (3a) and (3e), respectively.

(3a) It would be a real comfort to me ***if*** *you would make me feel we belonged to each other*.

(7) It would be a real comfort to me *for you* **to** *make me feel we belonged to each other*.

(3e) It seems ***as if*** *we're in a bad situation*.

(8) It seems ***that*** *we're in a bad situation*.

(d) Impossibility of replacement by equivalent adverbial links. Another clear indication that the subordinators in bold in (3) realize a complementizer function is their ability to be replaced by prototypical declarative complementizers (see (3e) and

(8) above), while they are not interchangeable with other adverbial links belonging to their original semantic domains. Note, for instance, that the conditional subordinator *on condition that* cannot substitute for *if* in (3a) and that *but that* cannot be replaced by the marker of exception *except that* in (3c).

(3a) It would be a real comfort to me ***if*** *you would make me feel we belonged to each other.*

(9) *It would be a real comfort to me ***on condition that*** *you would make me feel we belonged to each other.*

(3c) I don’t doubt ***but that/ that*** *she meant it*.

(10) *I don’t doubt ***except that*** *she meant it*.

(e) Coordination with prototypical complements. Furthermore, the subordinate clauses in (3) can be coordinated with prototypical complement clauses, thus testifying to their complement status. In (11), for instance, an *if*-clause is coordinated with a *that*-clause, both clauses functioning as complements of the complement-taking predicate *feel*.

(11) Now, driving the horse and sulky borrowed from Mynheer Schuyler, he felt ***as if*** *every bone was topped by burning oil* <u>*and*</u> ***that*** *every muscle was ready to dissolve into jelly and leave his big body helpless and unable to move*. (Brown K14)

(f) Pronominalization. Sub-clauses like those in (3) also meet another criterion for complementhood, namely pronominalization (see McCawley 1988: 143): as shown in (12) and (13), the clauses in (3) can be recovered by the anaphoric elements *that* and *so*, respectively, just like complements do.

(3a) It would be a real comfort to me ***if*** *you would make me feel we belonged to each other*.

(12) **That** would be a real comfort to me.

(3e) It seems ***as if*** *we’re in a bad situation*.

(13) It seems **so**.

(g) Pseudo-clefting. A final piece of evidence comes from pseudo-clefting: clauses introduced by minor declarative complementizers are co-referential with *what* in a pseudo-cleft construction, as shown by the comparison of (3a) and (14).

(3a) It would be a real comfort to me ***if*** *you would make me feel we belonged to each other*.

(14) What would be a real comfort to me would be ***if*** *you would make me feel we belonged to each other*.

The application of the aforementioned criteria clearly shows that the subordinate clauses introduced by *if*, *though*, *but*, *lest*, *as if* and *as though* in the examples in (3) should be analyzed as complements rather than as adjuncts, despite the fact that they are introduced by subordinators which typically function as adverbial connectives.

In our previous research, and taking as a starting point several historical and Present-day English corpora, we have shown that the adverbial function has been the original historical function for these links, while their complementizer use is a derived function. This is shown in Figure 1 (adapted from López-Couso and Méndez-Naya 2015: 191), which provides the time-depth of both the adverbial function (blue lines) and the complementizer use (red lines) of *if*, *though*, *but*, *lest*, *as if* and *as though*.

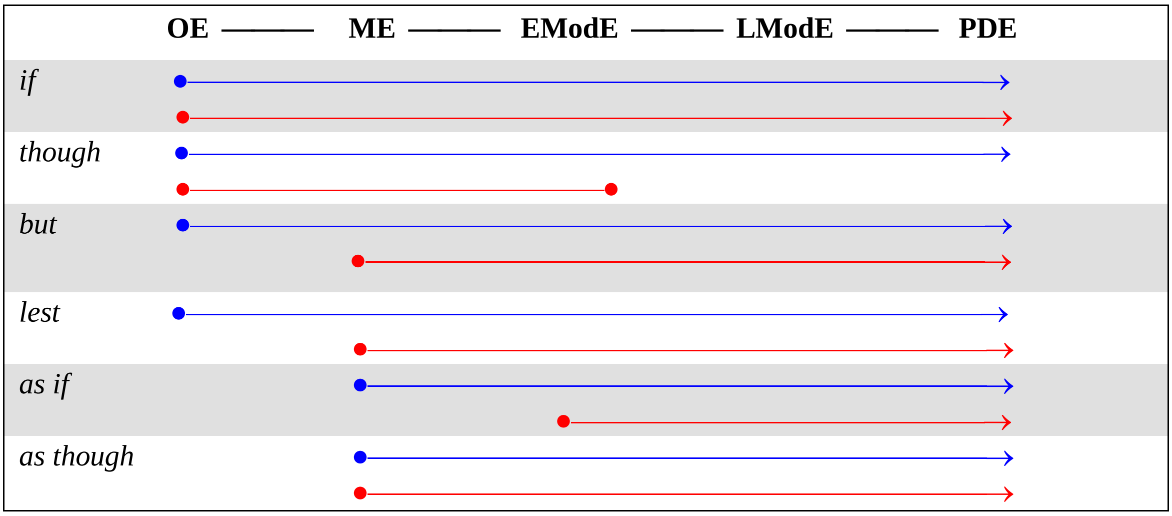


Figure 1: Timeline of *if*, *though*, *but*, *lest*, *as if* and *as though* in their functions as adverbial subordinators (blue lines) and declarative complementizers (red lines) (adapted from López-Couso and Méndez-Naya 2015: 191)

The adverbial function is not only the original use of these subordinators, but has also been the most frequent one all through their recorded history. Thus, for instance, the adverbial function of *lest* in the HC and ARCHER material represents almost 90% of all occurrences of the subordinator (see López-Couso 2007). Similarly, adverbial *as if* and *as though* show a ratio of 3:1 in these corpora with respect to their complementizing function (López-Couso and Méndez-Naya 2012b).

The low frequency of the complementizer use of these subordinators is one of the reasons why minor declarative complementizers have been overlooked in the literature, where the default interpretation for the italicized clauses in (3) is the adverbial one. In view of this, the way in which clauses introduced by these minor complement-clause links are parsed in corpora is worth examining.

4. The parsing of clauses introduced by minor declarative complementizers in the *Penn Parsed Corpora of Historical English*

As opposed to more traditional grammars like those by Quirk *et al*. (1985) and Biber *et al*. (1999), and in line with Huddleston and Pullum *et al*. (2002) and the Principles and Parameters framework (see Section 2 above), the annotators of the PPCHE establish a distinction between ‘complementizers’ and ‘prepositions’. In the first group they include *that*,[7] zero, *as* and Middle English *þe*, while the second group comprises both traditional prepositions and traditional subordinating conjunctions (other than ‘complementizers’).[8] In other words, traditional subordinating conjunctions are treated as prepositions taking a clausal complement (CP = ‘complementizer phrase’, that is, a clause headed by a complementizer) and are consequently tagged P.[9] As to the types of finite subordinate clauses (tagged CP), the parsing system distinguishes adverbial, *that*-clauses, degree complements, questions, exclamations and relative clauses.[10] Examples (15) and (16) below illustrate the tagging of an adverbial and a *that*-clause, respectively.

[7] Note that *that* is regarded as a clause subordinator, which can introduce complement clauses and also non-*wh*- relative clauses; see Huddleston and Pullum *et al*. (2002: 1034, 1056–1057).

[8] Huddleston and Pullum *et al*. (2002: 600) also include traditional adverbs within prepositions. They regard them as prepositions without a complement.

[9] Similarly, Huddleston and Pullum *et al*. (2002: 604) regard traditional subordinating conjunctions as prepositions taking “non-expandable [i.e. zero] content clauses” as complements.

[10] All subordinate clauses that are headed by a complementizer are labelled CP. In addition, there are four types of subordinate clauses whose primary label is IP (infinitives, small clauses, adjunct participials and absolutes). All subordinate clauses, both CP and IP, have a dash tag indicating their type. Finally, there are reduced relative clauses, which are labelled RRC. For further details, see https://www.ling.upenn.edu/hist-corpora/annotation/index.html.

(15) and *when the Ink was made Cleer again by the Oyl of Vitriol*, the affusion of dissolv'd <font> Sal Tartari <$$font> seem'd but to Praecipitate, (BOYLECOL-E3-P1,137.30)

```
( (IP-MAT (CONJ and)
         (PP (P when)
             (CP-ADV (C 0)
                    (IP-SUB (NP-SBJ-1 (D the) (N Ink))
                           (BED was)
                           (VAN made)
                           (IP-SMC (NP-SBJ *-1)
                                  (ADJP (ADJ Cleer))
                                  (ADVP (ADV again)))
                           (PP (P by)
                              (NP (D the)
                                  (N Oyl)
                                  (PP (P of)
                                     (NP (N Vitriol)))))))
         (, ,)
         (NP-SBJ (D the)
                (N affusion)
                (PP (P of)
                    (NP (VAN dissolv'd)
                      (LATIN (CODE <font>) (FW Sal) (FW Tartari) (CODE
                       <$$font>)))))
```

(16) I have already said, *the old general was kill'd by the shot of an arrow* (BEHN-E3-P1,155.118)

```
( (IP-MAT (NP-SBJ (PRO I))
         (HVP have)
         (ADVP-TMP (ADV already))
         (VBN said)
         (, ,)
         (CP-THT (CP-THT (C 0)
                      (IP-SUB (NP-SBJ (D the) (ADJ old) (ADJ general))
                             (BED was)
                             (VAN kill'd)
                             (PP (P by)
                                 (NP (D the)
                                    (N shot)
                                    (PP (P of)
                                        (NP (D an) (N arrow)))))
```

Let us now examine how structures involving minor declarative complementizers are parsed in the PPCHE. For practical purposes, we will focus on the Early Modern English period, when, as shown in Figure 1 above, the complementizer function is attested for all the items under study. By way of illustration, consider (17a), an example of a complement clause introduced by the minor declarative complementizer *if*, a complementizer which is typically associated with expressions meaning 'wonder' (López-Couso and Méndez-Naya 2001). This example will be compared with (17b), which features a parallel structure with the major complementizer zero.

(17a) Therefore it was no wonder ***if*** *we could not understand the Divine Essence* (BURNETROC-E3-P2,103.103)

(17b) […], tis noe wonder **Ø** *they should not like it*. (LOCKE-E3-P2,66.34)

According to our analysis, based on the criteria for complementhood outlined in Section 3 above, the two instances show an extraposed complement clause in subject function which is anticipated in the matrix clause by the dummy pronoun *it* (underlined in the examples). The only structural difference between the two instances lies in the choice of subordinator: the minor complementizer *if* in (17a) and the major complement-clause connective zero in (17b). Note that although the two clauses are structurally similar, complementizer selection clearly signals a semantic difference as regards the speaker's degree of commitment to the truth of the proposition expressed in the matrix clause, which seems to be higher in the case of (17b) than in (17a).

Here follows the parsing for these two sentences in the PPCEME:

```
(17a')  ( (IP-MAT (PP (ADV+P Therefore))
            (NP-SBJ (PRO it))
            (BED was)
            (NP-OB1 (Q no) (N wonder))
            (PP (P if)
                (CP-ADV (C 0)
                       (IP-SUB (NP-SBJ (PRO we))
                              (MD could)
                              (NEG not)
                              (VB understand)
                              (NP-OB1 (D the) (ADJ Divine) (N Essence)))))
          (. :))
```

```
(17b')   (NP-SBJ-1 (PRO $'t))
            (BEP $is)
            (CODE {TEXT:tis})
            (NP-OB1 (Q noe) (N wonder))
            (CP-THT-1 (C 0)
                     (IP-SUB (NP-SBJ (PRO they))
                            (MD should)
                            (NEG not)
                            (VB like)
                            (NP-OB1 (PRO it))))
            (. .))
```

As can be seen, even though the two examples contain parallel structures which only differ as regards the connective introducing the sub-clause (*if* in (17a) vs. zero in (17b)), the parsing fails to capture the obvious structural similarities which, in our view, exist between the two sequences. While in (17a) *if* is analyzed as a preposition P, taking a subordinate adverbial clause as complement (CP-ADV), in (17b) the sub-clause is tagged as a complement (CP-THT), attached immediately below the predicate (*be no wonder*).

```
(a)       (PP (P if)                        (b)      (CP-THT-1 (C 0)
              (CP-ADV (C 0)
```

We also find here a different treatment of the pronoun *it* in the two parsings: while in (17b) the zero clause is co-indexed with the anticipatory subject *it* (NP-SBJ-1), no co-indexing is present in the clause introduced by *if* in (17a) (NP-SBJ).

It should be noted, however, that the PPCEME is not completely consistent in the treatment of anticipatory *it* in such cases. In (18) below, an example of an *if*-clause complementing the matrix *it is meruayle*, very similar to the sequence in (17a), the *if*-clause is indeed co-indexed with the pronoun *it* in the matrix, just as the zero clause in our earlier instance (17b).

(18) and it is meruayle, ***if*** *thou scape with thy lyfe*, (FITZH-E1-H,101.376)

(18')

```
( (IP-MAT (CONJ and)
   (NP-SBJ-1 (PRO it))
   (BEP is)
   (NP-OB1 (N meruayle))
   (, ,)
   (PP-1 (P if)
        (CP-ADV (C 0)
               (IP-SUB (NP-SBJ (PRO thou))
                      (VBP scape)
                      (PP (P with)
                          (NP (PRO$ thy) (N lyfe))))))
   (. ,))
```

Similar analyses to the one provided for the *if*-clause in (17a) above are given in the PPCEME for clauses introduced by other minor declarative complementizers and for those featuring their *that*/zero counterparts. Examples (19a)–(19b) illustrate the use of *though*/*that* after a predicate meaning 'wonder' (López-Couso and Méndez-Naya 2001):

(19a) [...] meruayle it shall not be, ***thoughe*** *he be greued with pouertee*. (FITZH-E1-H,99.339)

(19a')

```
 (NP-OB1 (N meruayle))
   (NP-SBJ (PRO it))
   (MD shall)
   (NEG not)
   (BE be)
   (, ,)
   (PP (P thoughe)
       (CP-ADV (C 0)
              (IP-SUB (NP-SBJ-RSP (PRO he))
                     (BEP be)
                     (VAN greued)
                     (PP (P with)
                        (NP (N pouertee))))))
   (. .))
```

(19b) for it is maruell ***that*** *a sinner can without shame beholde this blessed Image?* (FISHER-E1-H,1,399.198)

(19b’)

```
( (IP-MAT (CONJ for)
        (NP-SBJ-1 (PRO it))
        (BEP is)
        (NP-OB1 (N maruell))
        (CP-THT-1 (C that)
                (IP-SUB (NP-SBJ (D a) (N sinner))
                       (MD can)
                       (PP (P without)
                          (NP (N shame)))
                       (VB beholde)
                       (NP-OB1 (D this) (VAN blessed) (N Image))))
        (. ?))
```

The same holds for *lest*, which is associated with predicates denoting fear (López-Couso 2007), as in (20a) vs. (20b):[11]

(20a) Did Cobham fear ***lest*** *you would betray him in Jersey?* (RALEIGH-E2-P1,1,218.132)

(20a’)

```
( (CP-QUE (IP-SUB (DOD Did)
               (NP-SBJ (CODE <font>) (NPR Cobham) (CODE <$$font>))
               (VB fear)
               (PP (P lest)
                   (CP-ADV (C 0)
                          (IP-SUB (NP-SBJ (PRO you))
                                 (MD would)
                                 (VB betray)
                                 (NP-OB1 (PRO him))
                                 (PP (P in)
                                    (NP (CODE <font>) (NPR Jersey) (. ?)
```

(20b) for he feared ***that*** *should he continew at Court*, (PERROTT-E2-H,33.12)

(20b’)

```
( (IP-MAT (CONJ for)
         (NP-SBJ (PRO he))
         (VBD feared)
         (CP-THT (C that)
               (IP-SUB (CP-ADV (IP-SUB (MD should)
                                   (NP-SBJ (PRO he))
                                   (VB continew)
                                   (PP (P at)
                                       (NP (N Court)))))
```

A similar parsing is given for *as if*-clauses dependent on propositional attitude predicates such as *seem* (López-Couso and Méndez-Naya 2012a, 2012b), as shown in (21a)–(21b).

[11] As noted by an anonymous reviewer, the variant *lest that*, as in (i) below, occurs in uncontroversial cases of complement clauses in the PCEEC, though not in the PPCEME.
(i) He ferythe ***lesse that*** *he schall neuer come home* (PASTON, I.656.9503)

Note that *as if* is not parsed as a unit in the PPCEME, but rather as two recursive prepositions.

(21a) so like our first parents before the fall, it seems ***as if*** *they had no wishes*, (BEHN-E3-P1,149.32)

(21a')

```
(NP-SBJ (PRO it))
(VBP seems)
(PP (P as)
    (PP (P if)
       (CP-ADV (C 0)
              (IP-SUB (NP-SBJ (PRO they))
                     (HVD had)
                     (NP-OB1 (Q no) (NS wishes))
```

(21b) and it seemes **Ø** *Sir Robert Bevell thinks our demaunds very unreasonable*. (MASHAM-E2-P1,103.59)

(21b')

```
( (IP-MAT (CONJ and)
        (NP-SBJ-1 (PRO it))
        (VBP seemes)
        (CP-THT-1 (C 0)
                 (IP-SUB (NP-SBJ (NPR Sir) (NPR Robert) (NPR Bevell))
                        (VBP thinks)
                        (IP-SMC (NP-SBJ (PRO$ our) (NS demaunds))
                               (ADJP (ADV very) (ADJ unreasonable)))))
        (. .))
```

The evidence provided for the (a) and the (b) sequences in (17) – (21) above so far indicates that parsing clearly masks crucial syntactic similarities between functionally parallel structures.

Another case in which syntactic structure seems to be masked by the parsing of the PPCEME concerns sequences such as (22), which involve insubordination, i.e. “the conventionalized main clause use of what, on *prima facie* grounds, appear to be formally subordinate clauses” (Evans 2007: 367). Note that in (22) the italicized clause is introduced by *as though*, but there is no clause in the context that can be claimed to be a matrix.[12]

(22) In that I am giltles? ***As though*** *they were gilty*. (MORERIC-E1-P1,37.179)

The parsing of (22) shows that the *as though*-clause is coded in the PPCEME as an adverbial subordinate clause:[13]

[12] We regard clauses of this kind as exclamatory clauses (López-Couso and Méndez-Naya 2012b: 324). See also Brinton (2014) on *as if*-exclamatory clauses.

[13] Note that the insubordinated clause in (22') is parsed as Fragment (FRAG): “FRAG should be thought of as a last resort for annotating material consisting of at least two constituents, for which there is not enough material to construct an IP” (https://www.ling.upenn.edu/hist-corpora/annotation).

(22’)

```
( (FRAG (PP (P In)
        (CP-ADV (C that)
             (IP-SUB (NP-SBJ (PRO I))
                  (BEP am)
                  (ADJP (ADJ giltles)))))
     (, ?)
     (PP (P As)
        (PP (P though)
          (CP-ADV (C 0)
               (IP-SUB (NP-SBJ (PRO they))
                    (BED were)
                    (ADJP (ADJ gilty))))))
     (. .))
```

The foregoing discussion has shown that clauses introduced by the minor declarative complementizers *if*, *though*, *lest*, *as if* and *as though* are invariably parsed as adverbial clauses in the PPCEME. There is, however, one case in which clauses headed by one of our ‘minor’ links are taken in the PPCHE, though not consistently, to be complements rather than adjuncts. This involves the complementizer *but* (*that*). In López-Couso and Méndez-Naya (1998), we identified two different sub-types of the connective: on the one hand, *but*$_1$, which is equivalent to ‘that not’, and is therefore a negative complementizer marking the sub-clause as negative, as shown in (23); on the other, *but*$_2$, as in (24), which means ‘that’, and typically occurs after negated predicates which are themselves inherently negative, such as *not doubt* or *not deny*. *But*$_1$, by contrast, is excluded from such contexts and occurs either with negated predicates (e.g. *not know*) or with inherently negative ones (e.g. *be a shame*).

(23) It is impossible ***but that*** *offences will come* (1582 Rhem; from López-Couso and Méndez-Naya 1998: 162) [i.e. ‘it is impossible that offences will not come’]

(24) Nor will any Man deny ***but that*** *every thing which is just, is good*; (1695, R. Preston, Cons. of Ph. 180; from López-Couso and Méndez-Naya 1998: 167) [i.e. ‘nor will any man deny that everything which is just, is good’]

In the annotation manual, *but* is one of those words which are treated individually (see Section 2 above), precisely due to its multifunctional nature. In the PPCHE parsing scheme, it is analyzed as a coordinating conjunction and tagged CONJ (e.g. *Jill laughed but Mary cried*), as a focus particle (FP; e.g. *It cannot be but a great folly*) and as a preposition (P; e.g. *Nobody but you*). The criterion used by the corpus annotators to distinguish between FP and P is whether *but* can be naturally replaced by *except* or *than*, in which case it is tagged as P:

> The distinction between the conjunction use of BUT on the one hand and the prepositional and focus particle uses on the other is generally clear, but the distinction between the latter two can be difficult. BUT is tagged as P if it can be replaced naturally by EXCEPT or THAN.

The manual also refers explicitly to cases of complementation with *but*, as follows:

> Inherently negative or questioning verbs (DENY, DOUBT, FEAR, HINDER, LET, MISTRUST, PREVENT, QUESTION) as well as other verbs or degree words when negated sometimes take finite clausal complements preceded by BUT. As in the NEG ... BUT construction, BUT is tagged FP and attached low (that is, as part of the complement clause).

An example of the complementizer use of *but*, more specifically *but*$_2$, meaning ‘that’, is given as (25a), where the *but*-clause is coded as CP-THT, just in the same way as the zero-clause in (25b).

(25a) However, I doubt not ***but** he is well*; (NHADD-1710-E3-P2,54.17)

(25a’)

```
( (IP-MAT (ADVP (WADV+ADV However))
         (, ,)
         (NP-SBJ (PRO I))
         (VBP doubt)
         (NEG not)
         (CP-THT (FP but)
               (C 0)
               (IP-SUB (NP-SBJ (PRO he))
                      (BEP is)
                      (ADJP (ADJ well))))
```

(25b) for I doubte not **Ø** *Sir <font> Thomas Wyat <$$font> hath bin examin’d of me, and hathe sayde what he could directly or indirectly*. (THROCKM-E1-H,I,68.C1.281)

(25b’)

```
( (IP-MAT-SPE (CONJ for)
             (NP-SBJ (PRO I))
             (VBP doubte)
             (NEG not)
             (CP-THT-SPE (C 0)
                      (IP-SUB-SPE (NP-SBJ (NPR Sir) (CODE <font>) (NP
                       Thomas) (NPR Wyat) (CODE <$$font>))
                                (HVP hath)
                                (BEN bin)
                                (VAN examin'd)
                                (PP (P of)
                                   (NP (PRO me))))
```

Interestingly, not all cases of *but*$_2$ are parsed in this way. Consider (26), a very similar example to (25a) above, but in which *but* is tagged as P taking a CP-ADV, rather than as CP-THT, even though the context makes it clear that *but* cannot be replaced by *except*.

(26) This Bishop is a temporall Lord, notwithstanding his sprituall title; and no doubt ***but** the flesh preuailes aboue the Spirit with him*; (JOTAYLOR-E2-P1,3,85.C2.281-282)

(26’)

```
( (CONJP (CONJ and)
        (NP (Q no)
            (N doubt)
            (PP (P but)
               (CP-ADV (C 0)
                     (IP-SUB (NP-SBJ (D the) (N flesh))
                            (VBP preuailes)
                            (PP (P aboue)
                                (NP (D the) (N Spirit)))
                            (PP (P with)
                                (NP (PRO him)))))))
        (. ;))
```

The same parsing is provided for (27), an example of *but*$_1$ ‘that not’, with the inherently negative predicate *be a great shame*.

(27) ‘Iff thys be trew,’ seyde Arthure, ‘hit were grete shame unto myne astate ***but that** he were myghtyly withstonde*.’ (CMMALORY-M4,45.1470)

(27’)

```
(NP-SBJ (PRO hit))
             (BED were)
             (NP-OB1 (ADJ grete)
                    (N shame)
                    (PP (P unto)
                       (NP (PRO$ myne) (N astate))))
             (PP (P but)
                (CP-ADV (C that)
                       (IP-SUB (NP-SBJ (PRO he))
                              (BED were)
                              (ADVP (ADV myghtyly))
                              (VAN withstonde))))
```

As seen, then, the parsing of the PPCHE recognizes that *but*-clauses can be complements, but this annotation is not always consistent throughout the corpora and does not account for the two uses of the complementizer *but*.

5. Concluding remarks

In this article we have discussed some of the issues that may derive from the syntactic annotation of corpora. In particular, we have examined the problems posed by the parsing of structures containing minor declarative complementizers in the PPCHE *vis-à-vis* canonical finite declarative complement clauses. We have shown that low-frequency phenomena such as the one considered in this article may go unnoticed, masked by annotation.

In our view, parsed corpora are undoubtedly useful for the analysis of highly-frequent and/or uncontroversial categories, but may overlook interesting features especially when dealing with low-frequency constructions, as shown here in connection with the annotation of minor declarative complementizers. While the parsing of the default finite complementation patterns with *that* and zero seems to be straightforward and therefore can be easily retrieved by means of the search engines, the annotation of complement clauses introduced by the minor complementizers discussed in this article is not completely devoid of problems. On the one hand, the parsing does not capture the obvious functional similarities between *that* or zero complement clauses and clauses headed by minor complementizers. On the other, the parsing of *if, though, lest, as if, as though* and *but* complement clauses is not always consistent, as shown in particular in the case of *but*$_1$ and *but*$_2$ and in the treatment of some anticipatory pronouns. Nevertheless, we believe that such minor weaknesses in the annotation of the Penn family of historical corpora do not at all diminish their value as indispensable tools for the study of the history of the English language.

References

ARCHER 3.1 = *A Representative Corpus of Historical English Registers* version 3.1. 1990–1993/2002/2007/2010/2013/2016. Originally compiled under the supervision of Douglas Biber and Edward Finegan at Northern Arizona University and University of Southern California; modified and expanded by subsequent members of a consortium of universities. Current member universities are Bamberg, Freiburg, Heidelberg, Helsinki, Lancaster, Leicester, Manchester, Michigan, Northern Arizona, Santiago de Compostela, Southern California, Trier, Uppsala, Zurich.

Bender, Emily and Dan Flickinger. 1999. Diachronic evidence for extended argument structure. In Gosse Bouma, Erhard Hinrichs, Geert-Jan M. Kruijff and Richard Oehrle eds. *Constraints and Resources in Natural Language Syntax and Semantics*. Stanford, CA: CSLI, 1–19.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Brinton, Laurel. 2014. The extremes of insubordination: Exclamatory *as if! Journal of English Linguistics* 42/2: 93–113.

Brook, Marisa. 2014. Comparative complementizers in Canadian English: Insights from early fiction. *University of Pennsylvania Working Papers in Linguistics* 20/2, Article 2.

Brook, Marisa. 2018. Taking it up a level: Copy-raising and cascaded tiers of morphosyntactic change. *Language Variation and Change* 30/2: 231–260.

Brown Corpus = A Standard Corpus of Present-day Edited American English, for use with Digital Computers (Brown). 1964, 1971, 1979. Compiled by W. N. Francis and H. Kučera. Brown University. Providence, Rhode Island. http://icame.uib.no/brown/bcm.html

Dancygier, Barbara and Eve Sweetser. 2005. *Mental Spaces in Grammar: Conditional Constructions*. Cambridge: Cambridge University Press.

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

Dirven, René. 1989. A cognitive perspective on complementation. In Dany Jaspers, Yvan Putseys, Wim Klooster and Pieter Seuren eds. *Sentential Complementation and the Lexicon: Studies in Honour of Wim de Geest*. Dordrecht: Foris, 113–139.

Elsness, Johan. 1982. *That* v. zero connective in English nominal clauses. *ICAME News* 6: 1–45.

Elsness, Johan. 1984. *That* or zero? A look at the choice of object clause connective in a corpus of American English. *English Studies* 65: 519–533.

Evans, Nicholas. 2007. Insubordination and its uses. In Irina Nikolaeva ed. *Finiteness: Theoretical and Empirical Foundations*. Oxford: Oxford University Press, 366–431.

Fanego, Teresa. 1990. Finite complement clauses in Shakespeare's English I & II. *Studia Neophilologica* 62: 3–21; 129–149.

Finegan, Edward and Douglas Biber. 1995. *That* and zero complementisers in Late Modern English: Exploring ARCHER from 1650–1990. In Bas Aarts and Charles F. Meyer eds. *The Verb in Contemporary English: Theory and Description*. Cambridge: Cambridge University Press, 241–257.

HC = *Helsinki Corpus of English Texts*. http://www.ota.ox.ac.uk/desc/1477

Heine, Bernd. 2002. On the role of context in grammaticalization. In Ilse Wischer and Gabriele Diewald eds. *New Reflections on Grammaticalization*. Amsterdam: John Benjamins, 83–101.

Huddleston, Rodney. 1971. *The Sentence in Written English: A Syntactic Study Based on an Analysis of Scientific Texts*. Cambridge: Cambridge University Press.

Huddleston, Rodney and Geoffrey K. Pullum *et al*. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Kaltenböck, Gunther. 2006. …*That* is the question: Complementizer omission in extraposed *that*-clauses. *English Language and Linguistics* 10/2: 371–396.

Kearns, Kate. 2007. Epistemic verbs and zero complementizer. *English Language and Linguistics* 11/3: 475–505.

Kortmann, Bernd. 1997. *Adverbial Subordination: A Typology and History of Adverbial Subordinators Based on European Languages*. Berlin: Mouton de Gruyter.

Kroch, Anthony and Ann Taylor. 2000. *The Penn-Helsinki Parsed Corpus of Middle English* (PPCME2). Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 4. http://www.ling.upenn.edu/ppche-release-2016/PPCME2-RELEASE-4

Kroch, Anthony, Beatrice Santorini and Lauren Delfs. 2004. *The Penn-Helsinki Parsed Corpus of Early Modern English* (PPCEME). Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, release 3. http://www.ling.upenn.edu/ppche-release-2016/PPCEME-RELEASE-3

Kroch, Anthony, Beatrice Santorini and Ariel Diertani. 2016. *The Penn Parsed Corpus of Modern British English* (PPCMBE2). Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 1. http://www.ling.upenn.edu/ppche-release-2016/PPCMBE2-RELEASE-1

Kytö, Merja. 1996. *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts* (third edition). Helsinki: Department of English, University of Helsinki.

Lakoff, Robin. 1968. *Abstract Syntax and Latin Complementation*. Cambridge, MA: The MIT Press.

López-Couso, María José. 1996. A look at *that*/zero variation in Restoration English. In Derek Britton, ed. *English Historical Linguistics 1994*. Amsterdam: John Benjamins, 271–286.

López-Couso, María José. 2007. Adverbial connectives within and beyond adverbial subordination: The history of *lest*. In Ursula Lenker and Anneli Meurman-Solin eds. *Connectives in the History of English*. Amsterdam: John Benjamins, 11–29.

López-Couso, María José and Belén Méndez-Naya. 1998. On minor declarative complementizers in the history of English: The case of *but*. In Jacek Fisiak and Marcin Krygier eds. *Advances in English Historical Linguistics*. Berlin: Mouton de Gruyter, 161–171.

López-Couso, María José and Belén Méndez-Naya. 2001. On the history of *if*- and *though*-links with declarative complement clauses. *English Language and Linguistics* 5/1: 93–107.

López-Couso, María José and Belén Méndez-Naya. 2012a. On the use of *as if*, *as though* and *like* in Present-day English complementation structures. *Journal of English Linguistics* 40/2: 172–195.

López-Couso, María José and Belén Méndez-Naya. 2012b. On the origin and development of comparative complementizers in English: Evidence from historical corpora. In Nila Vázquez ed. *Creation and Use of Historical English Corpora in Spain*. Newcastle upon Tyne: Cambridge Scholars Publishing, 311–333.

López-Couso, María José and Belén Méndez-Naya. 2014. The use of *if* as a declarative complementizer in English: Some theoretical and empirical considerations. In Alejandro Alcaraz-Sintes and Salvador Valera-Hernández eds. *Diachrony and Synchrony in English Corpus Linguistics*. Bern: Peter Lang, 85–107.

López-Couso, María José and Belén Méndez-Naya. 2015. Secondary grammaticalization in clause combining: From adverbial subordination to complementation in English. *Language Sciences* 47: 188–198.

Marcus, Mitchell, Beatrice Santorini and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The *Penn Treebank*. *Computational Linguistics* 19: 313–330. Reprinted in Susan Armstrong ed. *Using Large Corpora.* Cambridge, MA: The MIT Press, 273–290.

McCawley, James D. 1988. *The Syntactic Phenomena of English*. Chicago: The University of Chicago Press.

Mitchell, Bruce. 1985. *Old English Syntax*. 2 vols. Oxford: Clarendon Press.

Noonan, Michael. 1985. Complementation. In Timothy Shopen ed. *Language Typology and Syntactic Description*, Vol. 2. Cambridge: Cambridge University Press, 42–140.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Rissanen, Matti. 1991. On the history of *that*/zero as object clause links in English. In Karin Aijmer and Bengt Altenberg eds. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman, 272–289.

Rissanen, Matti. 2008. Corpus linguistics and historical linguistics. In Anke Lüdeling and Merja Kytö eds. *Corpus Linguistics: An International Handbook*, Vol. 1. Berlin: Mouton de Gruyter, 53–68.

Rooryck, Johan. 2000. *Configurations of Sentential Complementation: Perspectives from Romance Languages*. London: Routledge.

Santorini, Beatrice. 2016. *Annotation manual for the Penn Historical Corpora and the York-Helsinki Corpus of Early English Correspondence*. https://www.ling.upenn.edu/hist-corpora/annotation/index.html

Tagliamonte, Sali and Jennifer Smith. 2005. No momentary fancy! The zero ‘complementizer’ in English dialects. *English Language and Linguistics* 9/2: 289–309.

Taylor, John R. and Kam-Yiu Pang. 2008. Seeing as though. *English Language and Linguistics* 12/1: 103–139.

Torres Cacoullos, Rena and James A. Walker. 2008. On the persistence of grammar in discourse formulas: A variationist study of *that*. *Linguistics* 47: 1–43.

Warner, Anthony. 1982. *Complementation and the Methodology of Historical Syntax*. London: Croom Helm.

Corresponding author
María José López-Couso
Facultade de Filoloxía
Universidade de Santiago de Compostela
Av. de Castelao s/n
15782 Santiago de Compostela
Spain
e-mail: mljopez.couso@usc.es

received: February 2020
accepted: October 2020

RiCL Research in Corpus Linguistics

Review of Yáñez-Bouza, Nuria, Emma Moore, Linda Van Bergen and Willem B. Hollmann eds. 2019. *Categories, Constructions, and Change in English Syntax.* Cambridge: Cambridge University Press. ISBN: 978-1-108-41956-7. htps://doi.org/10.1017/9781108303576

Teresa Fanego
University of Santiago de Compostela / Spain

1. Introduction

This volume is part of the *Studies in English Language* series, edited by Merja Kytö for Cambridge University Press. In the introduction, Nuria Yáñez-Bouza, Emma Moore, Linda Van Bergen and Willem B. Hollmann write that the volume, "[w]hilst not a Festschrift" (p. 2), is a tribute to David Denison, Professor Emeritus of English Linguistics at the University of Manchester and former Smith Professor of English Language and Medieval Literature. Despite their denial, *Categories, Constructions, and Change in English Syntax* (henceforth CCChES) shares some of the features of memorial volumes: each of the four editors –as well as the assistant editor Ayumi Miura– was supervised by David Denison, and the contributors are friends and colleagues who have worked in close collaboration with Denison throughout his academic career.

The volume, however, differs from most Festschrifts in the quality of the individual contributions and their thematic coherence. Memorial volumes often consist of a wide range of essays of variable quality which, rather than being 'editorially integrated' (Reid 2009), lack thematic unity. In contrast, the essays collected here not only add substantially to our current knowledge of English syntax, but also engage deeply with the honouree's work by focusing on categorial and constructional description and change, two areas where Denison's research has left an enduring legacy.

DOI 10.32714/ricl.08.02.08

2. DISCUSSION

CCChES comprises an introduction ("Analysing English syntax past and present") by the four editors, plus fourteen chapters. The introduction opens with a touching and admirable summation of the significance of the honouree to both the editors and the field at large, and then clearly outlines the themes covered in the volume. This is structured into three parts: Part I is devoted to grammatical categories, Part II to constructions and constructional change, and Part III to comparative and typological approaches.

2.1. Part I: Approaches to grammatical categories and categorial change

Part I, comprising five chapters, addresses the fuzzy status of various grammatical categories, opening with John Payne's contribution "What is special about pronouns?" The focus here is on the use of personal pronouns as complements in an *of*-PP, as in *the brother of him*, an environment in which, according to Lyons's intuitive judgement (1986: 136), personal pronouns are ungrammatical or at least questionable, contrary to the alternative patterns with the *s*-genitive and the so-called oblique genitive, illustrated respectively in *his brother* and *that brother of his.* Lyons's observation provides the starting point for Payne's study; its goal is to provide a detailed empirical account of the *of*-PP construction based on late twentieth-century data retrieved from the *British National Corpus*. For this purpose, Payne extracted a random sample of 1,000 tokens of a search string consisting of any noun followed by *of* and a personal pronoun, of which, after manual filtering, 685 tokens were found to be instances of the *of*-PP construction under analysis. This confirms that personal pronouns can indeed occur as dependents in an *of*-PP, and can be employed in a wide range of semantic relations, with quantity (185 tokens), as in *a rare* ***lot*** *of them*, and theme (164 tokens), where the head noun is a nominalisation corresponding to a transitive verb, as in *the critical* ***evaluation*** *of it*, predominating. Personal pronouns as *of*-dependents are not excluded even with semantic relations which are known to be quite strongly predisposed towards the *s*-genitive (see Rosenbach 2002), as is the case with kinship, attested three times in Payne's random sample and illustrated in (1) below.

(1) And **the father of them all** David Smith is represented by a selection of sculptures all this month. (EBT 2626)

On the other hand, a few of the semantic relations found with *of*-PPs are not available to the *s*-genitive construction, such as quantity (*the rest of you*/**your rest*) or content (*the idea of it*/**its idea*), among others. The quantitative and qualitative results thus confirm the claim made in earlier research by Payne (e.g. Payne and Huddleston 2002: 476–478; Payne and Berlage 2014) that the set of semantic relations available to the *of*-PP construction is a superset of those available to the *s*-genitive, so that a speaker's knowledge must also include the knowledge of the different variables that predispose towards one construction or the other.

Like Payne, Bas Aarts deals in Chapter 2 ("What *for*?") with Present-Day English (PDE) data. He revisits an issue briefly discussed in Aarts (2007: 219–222), namely the categorial status of *for* in sequences such as those in (2)–(3):

(2) *For* + NP: Hold it **for a moment**. (S1A-002 127)

(3) *For* + [NP *to* VP]: **For the roles to be reversed** would be a tragedy for many Conservative MPs and voters. (W2E-004 064)

Example (2) involves the use of *for* as a preposition governing a NP complement; (3), on the other hand, represents a usage in which *for* is most commonly analysed as a subordinator introducing an infinitival clause. *For* NP *to*-infinitivals have received considerable attention in the literature (Fischer 1988; Huddleston and Pullum 2002: 1181–1183; De Smet 2009, among many others), the usual assumption being that they go back to structures such as (4), with an 'organic' *for*-NP dependent on an element outside the *to*-infinitive clause. Over the course of time the NP would have been reanalysed as part of the *to*-infinitive clause with which it happened to co-occur. This "meant that the preposition *for* lost whatever meaning it had and became an 'inorganic' infinitival subject marker" (De Smet 2009: 1743), enabling the extension of the construction to radically new environments, as in (5), where the new analysis of *for* as a subordinator is the only possible option (examples from Aarts 2019: 58–59):

(4) It is good [$_{\text{PP}}$ for a man] [not to touch a woman].

(5) It is a rare thing [for [a night to pass without one or other of us having to trudge off]].

Aarts offers a detailed account of the guises of *for* in a wide array of constructions and critically reviews the literature. He concludes by arguing that there are strong reasons in favour of categorising *for* as a preposition in all of its uses, an analysis that does away "with the overly complicated historical account in which *for* ceases to be a preposition and becomes a subordinator" and "allows for a parsimonious and elegant parallel way" (p. 76) of treating the various kinds of constructions where *for* is used in PDE, such as those exemplified above, or cases in which *for* is followed by an *-ing* clause, as in (6); this is an environment where *for*, at any rate, is usually considered a preposition, since *-ing* clauses, unlike infinitives, share most features of NPs, including the ability to occur as prepositional complements (e.g., *On hearing the news, she telephoned her father*):

(6) She hated herself **for allowing the policeman to intimidate her**. (W2F-009 100)

Some of Aarts's arguments in favour of a unified analysis are persuasive, but his proposal seems to leave a few loose ends. *For* NP *to*-infinitivals, for instance, can freely occur as preverbal (3) and extraposed subjects (5), but these two slots, by contrast, are not available either to *for* NPs or to *for -ing* clauses, as Aarts himself acknowledges (pp. 67; 75, footnote 14). If *for*, as he contends, is always a member of the category of prepositions, one would expect it to have the same distribution across all clause slots, whether it is followed by a noun phrase, an *-ing* form, or a *to*-infinitive.

Dan McColm and Graeme Trousdale turn to the recent history of English in "Whatever happened to *whatever*?" Their focus is on the use of *whatever* as a discourse marker (DM), as in (7):

(7) All right. **Whatever**. I'll let Rush speak for millions and myself. (COCA, 2012; McColm and Trousdale 2019: 84)

Brinton (2017: 268–283) suggests that this usage is a late twentieth-century phenomenon, with two potential syntactic origins: the use of *whatever* as a general extender in a coordinate structure (e.g., *He wants to be a film star* ***or whatever***), and the chunk *whatever you* V, where *whatever* is followed by a second-person subject and a verb of cognition, volition, or speaking (e.g., *Whatever you please*). The latter type, Brinton argues, is the more likely source, because its discourse context is the same as that of the DM, namely dialogic and associated with a certain kind of speaker/writer stance, often irritability or exasperation. McColm and Trousdale complement the qualitative analysis of the development of *whatever* presented by Brinton (2017) with a quantitative study based on large random samples extracted from several synchronic and diachronic corpora. The results serve to confirm Brinton's hypothesis that chunks of the form *whatever you* V "have a significant role to play in the development of the discourse marker *whatever*, which typically serves as a distinct conversational turn" (p.

104). But in addition, McColm and Trousdale suggest that contexts involving the general extender have also contributed to the development of the discourse marker, since these –like the chunk *whatever you* V itself– often carry a certain attitude or stance, particularly one which appears to be dismissive of the addressee, as is the case in (8):

(8) Well, Willoughby Pastures, –or **whatever your name is**–, you'll get yourself into the papers this time. (COHA, 1877; McColm and Trousdale 2019: 97)

In light of the above, they conclude that the diachrony of the DM *whatever* can be understood as an example of 'bolstering': while one construction may be the most likely source of a new form-meaning pairing, other constructions serve to strengthen the representation of the new pattern, "bolster[ing] it via a formal or functional alignment (or both)" (p. 81). One could point out here that the coining of a new label was perhaps not strictly necessary, since the notion of bolstering seems to be analogous to Van de Velde's (2014: 147) 'horizontal construction links'; these are also based on similarities in the form and/or meaning pole, and have been shown to play an important role in the synchronic network of constructions as well as in diachronic change (see further Hoffmann 2018).

In the chapter "Are comparative modals converging or diverging in English? Different answers from the perspectives of grammaticalisation and constructionalisation," Elizabeth Closs Traugott addresses the history of the comparative modals *better*, *rather* and *sooner* from the perspective of the construction grammar formalism laid out by Traugott and Trousdale (2013). In PDE, comparative modals differ semantically in that *rather* and *sooner* code preference, while *better* expresses advice. Traugott's goal is to revisit a topic inspired by Denison and Cort's research (2010) on the rise of 'bare' *better* (e.g., *You better go*), and especially to complement Van linden's (2015) study on the development of the three comparative modals in recent American English. For this purpose, Traugott traces their history in British English, using the *Middle English Dictionary* and several corpora of Early and Late Modern English. She concludes that *rather* and *sooner* emerged as preference modals by the sixteenth century; see (9):

(9) Yett **haid** I **rether** dye for his sake.
'Yet I would rather die for his sake' (¿c1500 *Grevus Ys* (Sln 1584) 87; Traugott 2019: 114)

For *better* she identifies sporadic "preference readings" (p. 119) in the seventeenth century, and entrenchment as a modal auxiliary by the early eighteenth century (pp. 118, 126–127). She also finds that at this stage *better* had already specialised in its current advisory meaning, which is the only one attested in the 975 instances of *had better* recorded in her data from the *Old Bailey Corpus* (1720–1913). The evidence discussed by Traugott is rich and varied, but her chronology of the changes, which suggests a rather late and abrupt emergence of *better* as an auxiliary, can now be revised thanks to the availability of big corpora such as EEBO BYU (1470s–1690s; see Davies 2017). EEBO shows clearly that by the sixteenth century *better* was already well established both as a preference modal (10) and as an advice modal (11), this latter usage arising naturally out of the preference usage.

(10) VXOR: what doest thou here in this countree, me thinke thou art a scot by thy tongue. MENDICUS: trowe me […], i **had better** bee hanged in a withie of a cowtaile, then be a rowfooted Scotte, for thei are euer sare and fase: (EEBO 1564 William Bullein, *A dialogue bothe pleasaunte and pietifull*)

(11) now the time doth not serue any longer to geue men brickbattes for turfes, or to make them beleeue that the Moone is made of greene cheese: for euerie one will pretend now to know how the world walkes: therefore he **had better** haue held his tongue touching this matter: (EEBO 1579 Marnix van St. Aldegonde/John Stell/George Gilpin, *The bee hiue of the Romishe Church*)

The final chapter in Part I, "The definite article in Old English: Evidence from Ælfric's *Grammar*," is by Cynthia L. Allen, who addresses the question whether a category of definite article already existed in Old English (OE). Studies that take the point of view that OE had no definite article, or at least that definiteness marking was not obligatory, are numerous (e.g., van Gelderen 2007: 297; Watanabe 2009; Sommerer 2015: 112). Authors such as Crisma (2011), however, adopt a different position and argue that in prose writing, subject and object NPs, that is, referential NPs in argument function, were already regularly marked for definiteness in late OE. Allen's study, which is inspired by Crisma (2011), therefore focuses on subjects and objects; predicative NPs (e.g., *he wæs to* ***cyninge*** *gecoren* 'he was chosen as **king**'), which are non-referential, lack definiteness marking even in PDE, and thus do not constitute good evidence. As a source of data, Allen employs Ælfric's *Grammar* of Latin, an adaptation of the *Excerptiones de Prisciano*. Latin is a language without a category of definite article, and this allows Allen to show that in the English translations of Latin sentences Ælfric consistently adds the relevant form of *se* whenever a definite interpretation of the original would be the most likely one. Her meticulous philological study thus confirms Crisma's claim (2011) that English had a definite article prior to the early Middle English (ME) stage that is most commonly accepted as the period when this category emerged. This finding is in agreement with the results independently arrived at by Sommerer (2018), on the basis of a quantitative analysis of the *Parker* and *Peterborough* chronicles; like Crisma (2011) and Allen (2019), Sommerer (2018: 300, 312) concludes that at some point between early and late OE definiteness marking became obligatory in all referential cases.

2.2. Part II: Approaches to constructions and constructional change

Like Part I, Part II consists of five chapters, and opens with Bettelou Los's study on "How patterns spread: The *to*-infinitival complement as a case of diffusional change, or '*to*-infinitives, and beyond!'." Los revisits her earlier work on *to*-infinitives (Los 2005) in the light of new insights about the spread of the gerund as a verb complement provided by De Smet (2013). Her goal is to investigate how De Smet's model of analogical change can account for the diffusion of *to*-infinitival complements in the early stages of English. She proposes the recognition of five developmental stages: Stage I involves verbs of spatial manipulation with meanings like PDE *force*. Stage II pertains to verbs such as OE *ontendan* 'kindle, set fire to', which extended their meanings metaphorically to 'fire someone up, inspire someone to do something'. In Stage III the *to*-infinitive spread to verbs with a similar directive meaning, namely the verbs of commanding and permitting. Importantly, this extension allowed the *to*-infinitive to acquire a more abstract meaning, similar to that of a subjunctive clause. The subjunctive clause "may have provided a new model, so that the *to*-infinitive started to appear with verbs that […] had no directive meaning: they were verbs of intention with meanings of intending, hoping, trying, promising" (pp. 163–164). This is Los's Stage IV, which witnesses the diffusion of *to*-infinitives to verbs such as OE *giernan* 'yearn', *secan* 'seek' or *swerian* 'swear'. Towards the late fourteenth century, *to*-infinitives also became available with verbs of thinking and declaring, such as *believe*, *profess*, *say*, *think* and the like, in the so-called 'Exceptional Case-Marking' (ECM) construction, as in (12) below:

(12) *þat man [...]* ***is seid*** *to have an heed*
'the man is said to have a head'
(c1390, *Wyclifite Sermons*; quoted from Warner 1982: 136)

This is Stage V, which clearly "cannot be made part of any natural progression from the previous stages" (p. 168). On the basis of work by Dreschler (2015), Los suggests that this time the model for extension was an adjectival or participial construction with *to*-infinitival postmodification, as in (13), rather than a verbal construction:

(13) *&* ***wes iwunet*** *ofte to cumen wið him to his in.*
'and was wont often to come to him to his lodgings'
(c1225(1200), cmjulia.96.12; Dreschler 2015: 176)

This type of pattern helps to make sense of the fact that from their earliest emergence in the ECM construction, *to*-infinitives occur frequently in passive sentences, such as (12) above. Following Dreschler (2015: 176–177), Los argues that the adjectival/participial construction provided a template for the emergence of ECM-passives. She acknowledges, however, that the availability of *to*-infinitives with verbs of thinking and declaring remains "the odd one out in the scenario" (p. 168) of diffusional change from one class of verbs to the next which her study envisages.

Ayumi Miura's chapter "*Me liketh*/*lotheth* but *I loue*/*hate*: Impersonal/non-impersonal boundaries in Old and Middle English" addresses impersonal constructions, one of the most extensively researched topics in English historical syntax (e.g., van der Gaaf 1904; Elmer 1981; Denison 1990, 1993: 61–102; Allen 1995; Möhlig-Falke 2012; Light and Wallenberg 2015; Miura 2015; Castro-Chao 2019, among many others). Miura (2015) investigated, with reference to ME, the range of factors determining the use of the verbs *like* and *loathe* as impersonal, as opposed to the use of *love* and *hate* as non-impersonal. In the present analysis Miura examines whether the generalisations made in her earlier study can be extended to the OE period, and to the near-synonymous phrasal impersonals *be*/*have lief* and *be loath*, which are usually neglected in the literature on impersonals, as pointed out by Denison (1990: 125). She shows, with data from several corpora, that causation is the most important factor for drawing the boundary between impersonal and non-impersonal predicates. The verbs *like* and *loathe* as well as the phrasal impersonals *be lief*/*loath* are all attested in both impersonal and causative constructions in OE and ME, whereas the non-impersonal verbs *love* and *hate* have apparently never been causative in their history. (14) is an example of the causative use of *be loath*, with the Cause argument appearing as nominative subject:

(14)	*seo ceorung*	***is***	*swyðe* ***lað***	*Gode*
	the murmuring	is	very loath	God-DAT

‘the murmuring is very disgusting to God’ (coaelive,ÆLS_[Pr_Moses]; quoted from Miura 2019: 181)

As regards *have lief*, which emerged in ME as a new phrasal impersonal, it is not attested in causative use, contrary to expectations, but Miura suggests, quite plausibly, that analogy with *be lief* “may have provided sufficient motivation for its impersonal use” (p. 189).

Laurel J. Brinton’s chapter (“*That’s luck, if you ask me*: The rise of an intersubjective comment clause”) moves on to pragmatics, in a study that nicely ties in with McColm and Trousdale’s analysis in Part I of the DM *whatever.* In previous work Brinton investigated the development into comment clauses of *if*-conditionals such as *if you will* (Brinton 2008), *if you choose*/*like*/*prefer*/*want*/*wish* (Brinton 2014), and *if I may say so* (Brinton 2017). In the present chapter she traces the related development of *if you ask me* from having a literal meaning in the protasis of a direct condition (*If you ask me, I’m required to give it*) to its use as a politeness marker attached to an expression of opinion or evaluation by the speaker (*Well, it is the trick of the trade, if you ask me*). Examples of *if you ask me* serving such a function are not attested until the late nineteenth century, but other members of the network of pragmaticalised *if*-conditionals examined by Brinton, such as *if I may say so*, appear fully formed as early as the sixteenth century. According to Brinton, the fact that in the history of English *if*-clauses repeatedly exhibit this process of change from content to procedural meaning, and from nonsubjective to (inter)subjective meaning, calls “for a better understanding of the construction in general” (p. 209). Of relevance here is recent work by Lastres-López (2020a; also 2020b: 50), who proposes a pragmaticalisation cline for *if*-clauses, based on data from English spoken discourse.

We turn now to Sylvia Adamson’s contribution “Misreading and language change: A foray into qualitative historical linguistics,” whose goal is “readjust[ing] the balance between quantitative and qualitative approaches” (p. 212) to the history of English. As a case study, she focuses on the relative pronoun system, which was subject to significant variability in the Early Modern English (EModE) period, prior to its regularisation during Late Modern English (LModE). She discusses at length the reactions from eighteenth-century grammarians to a passage in Shakespeare’s *2 Henry VI* 3.2.161–165 where the relative *who* has a nonpersonal noun as antecedent (i.e., *the* *labouring heart, /Who […]*), a usage which was common at the time (Fanego 2016: 188). Adamson's concluding observation that "the challenge for future researchers is to determine how far qualitative analysis can be methodised" (p. 233) is one which will appeal to all readers.

The last chapter in Part II ("The conjunction *and* in phrasal and clausal structures in the *Old Bailey Corpus*) is by Merja Kytö and Erik Smitterberg, who look at *and*-coordination in trial proceedings, as represented in two different periods (1753–1785; 1850–1881) of the *Old Bailey Corpus.* Their starting point is the finding in Biber *et al.* (1999: 81) that in PDE conversation, *and* tends to be a clause-level connector, while the opposite holds true for academic prose, where *and* is more typically used at the phrase level. Within these two categories of coordination, clausal and phrasal, Kytö and Smitterberg also include what they label *V and V* coordination, which they consider to be a subtype of the clausal uses. *V and V* coordination –more commonly referred to in the literature as 'pseudo-coordination' (e.g., Quirk *et al.* 1985: 978–979)– consists of "two movement verbs conjoined by *and* in a set pattern that […] could be understood to form one entity of action" (p. 240), for instance, in *I went and enquired in the places.* The data reveal two clear diachronic trends, both affecting "the two patterns that seem characteristic of orality in PDE" (p. 247), namely clausal coordination and *V and V* coordination. Clausal coordination becomes more frequent over time, a result which the authors interpret as indicating that trial proceedings may be incorporating rising numbers of oral features, as part of the process of colloquialisation (Mair 1997: 202–205) documented in other written genres in the modern period. The *V and V* pattern, in contrast, becomes much less frequent diachronically, so that the authors hypothesise "that such constructions were felt not to be suitable for a formal courtroom setting and thus increasingly avoided" (p. 247); this suggestion, however, is at odds with the fact that trial proceedings seem to have become more oral and colloquial, to judge from the growth in frequency of clausal coordination mentioned above. A search in the *Old Bailey Corpus* for V + infinitive sequences of the type *I **went see** one of the teachers*, *I **go get** the paper every morning*, etc. (see further Flach 2015) might have thrown light on the development of the *V and V* pattern itself: diachronic work by Bachmann (2013) and Ross (2018) shows that during LModE *V and V* coordinations steadily lost ground to V + infinitive combinations, as part of a process of increasing auxiliation.

2.3. Part III: Comparative and typological approaches

Part III comprises four chapters focusing on the comparison of British English with other varieties of English, and with Germanic and Romance languages. The first chapter, by Olga Fischer and Hella Olbertz, discusses "The role played by analogy in processes of language change: The case of English *have-to* compared to Spanish *tener-que.*" It offers an analysis of the development of the semi-modal *have-to*, which is compared to the Spanish construction with *tener-que*. The obligative semi-modal *have-to* is usually assumed to have gone through a slow grammaticalisation process involving various developmental stages (Krug 2000: 55–56): from a possession schema (*I have a letter*) to a possession schema + purpose/goal adjunct (*I have a letter to write*) to a final stage where *have-to* functions as a unit expressing the modal notion of obligation (*I have to write a letter*). In an earlier account of the origins of *have-to*, Fischer (1994) saw the word order change –whereby *have* and the *to*-infinitive became adjacent due to increased SVO order over the course of the ME period– as the only cause for the changes in *have-to*. In this chapter, as already noted in Fischer (2015), it is argued instead that the new construction with *have-to* was supported analogically by other ME constructions expressing necessity, notably constructions involving the verb NEED (e.g., *Me* ***nedith*** *not no lenger doon…* 'It is no longer necessary for me to do…') and verbo-nominal combinations of *have*, *be* and *must* with the noun *need* (e.g., *þei* ***had nede*** *to ride in þat contrey* 'they had a need to ride in that country'). These neighbouring constructions "all contributed to the 'necessity' meaning that *have-to* acquired […], a development that the traditional gradual semantic-pragmatic grammaticalisation account cannot really explain" (p. 260). Two studies directly relevant to the analysis of *have-to* presented in this chapter, but inexplicably not mentioned by the authors, are those by Loureiro-Porto (2009, 2010). These exhaustively trace the development, from early OE to the eighteenth century, of both the verb *need* and the synonymous verbo-nominal constructions *be/have need* and *be/have tharf*; *be/have tharf* combinations, which predate the phrasal patterns with the noun *need*, are very frequently attested (205 tokens) in Loureiro-Porto's data. The second part of Fischer and Olbertz's chapter presents the development of the Spanish modal construction *tener-que*, currently the most popular expression of necessity in Spanish; as in the case of English *have-to*, neighbouring possession-based periphrases (e.g., *haber/aver-de* 'have to') appear to have played an analogical role. In its emphasis on the importance of multiple sources in the development of new constructions, this chapter is thus a nice follow-up to McColm and Trousdale's analysis, earlier in the volume, of the DM *whatever* as emerging out of several patterns that bolstered the new pattern via a formal or functional alignment.

In "Modelling step change: The history of *will*-verbs in Germanic" Kersti Börjars and Nigel Vincent look at the development of *will*-verbs on the basis of evidence from English, Danish, Dutch, Icelandic, and Swedish. All these languages have *will*-verbs that can be traced back to the Proto-Indo-European root **wel-* 'want, desire'. The chapter opens with a detailed description of the form and structure of the different *will*-verbs and of the categorial properties of their complements. After this the authors move on to meaning; in order to compare the semantics of *will* in the languages investigated, they use as a starting point the grammaticalisation cline in Bybee *et al.* (1994: 256), which envisages a development from Desire > Willingness > Intention > Prediction. In light of the evidence examined, Börjars and Vincent propose a reconceptualisation of the semantic connections as a three-stage cline of Desire > Intention > Prediction, with a "bifurcating diachronic route" (p. 302) from Desire to Willingness; this mirrors the fact that willingness is a meaning that naturally arises from any WANT verb, for instance English *want* (e.g., *Do you want to pass me the lunch menu?*), without necessarily developing further along the grammaticalisation path mentioned above. The observed changes are modelled within the theoretical framework of Lexical-Functional Grammar (Börjars and Vincent 2017).

Benedikt Heller and Benedikt Szmrecsanyi report on "Possessives world-wide: Genitive variation in varieties of English." This study, which complements Payne's study on genitive variation in Part I, combines the assumptions of Probabilistic Grammar (e.g., Bresnan 2007) with scholarship on World Englishes. The study addresses two questions: (a) the extent to which varieties of English have different grammars for genitive choice; and (b) what probabilistic constraints tend to make a difference across the varieties. Two genitive variants, the *s*-genitive and the *of*-genitive, are examined in nine different varieties of English from around the world: four Inner Circle varieties (Canada, Great Britain, Ireland, New Zealand), two advanced Outer Circle varieties (Jamaica, Singapore; see Schneider 2007) and three other Outer Circle varieties (Hong-Kong, India, Philippines). A key finding is that the *s*-genitive is more frequent in Inner Circle varieties than in L2 varieties of the Outer Circle. This difference is attributed to the well-known fact that contact varieties avoid synthetic structures, so the hostility in Outer Circle varieties towards clitic *-s* may have to do with the mode of language acquisition in these varieties. As regards question (b) above, the varieties under scrutiny are found to fall into two groups: in Group 1 (British English, Indian English, Jamaican English, and Philippine English) possessor length appears to be the most important language-internal factor in genitive choice; in Group 2 (Canadian English, Hong-Kong English, Irish English, New Zealand English, and Singapore English) possessor animacy is the top-ranked constraint.

The final chapter in Part III and in the volume is also concerned with varieties of English. In "American English: No written standard before the Twentieth Century?" Christian Mair takes as a starting point Schneider's Dynamic Model (2007) for the emergence of new varieties of English. According to Schneider, American English is the only postcolonial variety which has fully completed the five stages of emancipation from British English which the Dynamic Model postulates. More specifically, the Spanish-American War of 1898, which resulted from a new sense of national self-confidence in the USA and signalled a growing willingness to play a role on the world stage, is taken by Schneider (2007: 291) as the boundary between Phase 4 (endonormative stabilisation) and Phase 5 (differentiation) of American English, with this latter phase thus covering the twentieth and twenty-first centuries. Mair demonstrates, however, that Schneider's chronology of the emancipation of American English should be modified, and that the clear and consistent differentiation of British and American written standards has in fact to be placed "well into the twentieth century (and is in several instances still going on today)" (p. 337). This conclusion is supported by extensive evidence drawn from large and small corpora, and from linguistic features at the levels of orthography, morpholexis, and syntax. These include, among others, *-or*/ *-our* (e.g., *color*/*colour*) and *-er*/*re* (e.g., *centre*/*centre*) spellings, the morpholexical variants *toward*/*towards* and *gotten*/*got*, and the complementation patterns of the verbs *help* and *prevent.* The chapter combines detailed philological analysis of the individual examples with the statistical profiling of large masses of text, an integrative approach in which David Denison has always excelled.

As I mentioned at the beginning, and despite the minor points I have raised throughout these pages, this carefully edited volume is an outstanding collection of papers that address major issues in the field of English syntax from both a synchronic and diachronic perspective, and from a variety of methodological orientations, theoretical and applied. As such, it will no doubt attract the large readership it deserves.

References

Aarts, Bas. 2007. *Syntactic Gradience: The Nature of Grammatical Indeterminacy.* Oxford: Oxford University Press.

Allen, Cynthia L. 1995. *Case Marking and Reanalysis: Grammatical Relations from Old to Early Modern English.* Oxford: Clarendon Press.

Bachmann, Ingo. 2013. Has *go*-V ousted *go-and*-V? A study of the diachronic development of both constructions in American English. In Hilde Hasselgård, Jarle Ebeling and Signe Oksefjell Ebeling eds. *Corpus Perspectives on Patterns of Lexis*. Amsterdam: John Benjamins, 91–112.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English.* Harlow, Essex: Pearson Education Limited.

Börjars, Kersti and Nigel Vincent. 2017. Lexical-Functional Grammar. In Adam Ledgeway and Ian Roberts eds. *The Cambridge Handbook of Historical Syntax*. Cambridge: Cambridge University Press, 642–663.

Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston and Wolfgang Sternefeld eds. *Roots: Linguistics in Search of Its Evidential Base.* Berlin: Mouton de Gruyter, 75–96.

Brinton, Laurel J. 2008. *The Comment Clause in English: Syntactic Origins and Pragmatic Development.* Cambridge: Cambridge University Press.

Brinton, Laurel J. 2014. *If you choose/like/prefer/want/wish*: The origin of metalinguistic and politeness functions. In Marianne Hundt ed. *Late Modern English Syntax*. Cambridge: Cambridge University Press, 271–290.

Brinton, Laurel J. 2017. *The Evolution of Pragmatic Markers in English. Pathways of Change.* Cambridge: Cambridge University Press.

Bybee, Joan, Revere Perkins and William Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World.* Chicago and London: The University of Chicago Press.

Castro-Chao, Noelia. 2019. Changes in argument structure in Early Modern English with special reference to verbs of desire: A case study of *lust*. *Research in Corpus Linguistics* 7: 129–154.

Crisma, Paola. 2011. The emergence of the definite article in English: A contact-induced change? In Petra Sleeman and Harry Perridon eds. *The Noun Phrase in Romance and Germanic: Structure, Variation, and Change*. Amsterdam: John Benjamins, 175–192.

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

De Smet, Hendrik. 2009. Analysing reanalysis. *Lingua* 119/11: 1728–1755.

De Smet, Hendrik. 2013. *Spreading Patterns. Diffusional Change in the English System of Complementation.* Oxford: Oxford University Press.

Denison, David. 1990. The Old English impersonals revived. In Sylvia Adamson, Vivien A. Law, Nigel Vincent and Susan Wright eds. *Papers from the 5th International Conference on English Historical Linguistics: Cambridge, 6–9 April 1987*. Amsterdam: John Benjamins, 111–140.

Denison, David. 1993. *English Historical Syntax: Verbal Constructions*. London: Longman.

Denison, David and Alison Cort. 2010. *Better* as a verb. In Kristin Davidse, Lieven Vandelanotte and Hubert Cuyckens eds. *Subjectification, Intersubjectification and Grammaticalization.* Berlin: Mouton de Gruyter, 349–383.

Dreschler, Gea. 2015. *Passives and the Loss of Verb Second: A Study of Syntactic and Information-Structural Factors.* Utrecht: LOT Publications.

Elmer, Willy. 1981. *Diachronic Grammar: The History of Old and Middle English Subjectless Constructions*. Tübingen: Niemeyer.

Fanego, Teresa. 2016. Shakespeare's grammar. In Bruce R. Smith ed. *The Cambridge Guide to the Worlds of Shakespeare, Volume 1: Shakespeare's World 1500–1660.* Cambridge: Cambridge University Press, 184–191.

Fischer, Olga. 1988. The rise of the *for NP to V* construction: An explanation. In Graham Nixon and John Honey eds. *An Historic Tongue: Studies in English Linguistics in Memory of Barbara Strang.* London: Routledge, 67–88.

Fischer, Olga. 1994. The development of quasi-auxiliaries in English and changes in word order. *Neophilologus* 78/1: 137–164.

Fischer, Olga. 2015. The influence of the grammatical system and analogy in processes of language change: The case of the auxiliation of have-*to* once again. In Fabienne Toupin and Brian Lowrey eds. *Studies in Linguistic Variation and Change: From Old to Middle English.* Newcastle upon Tyne: Cambridge Scholars Publishing, 120–150.

Flach, Susanne. 2015. *Let's go look at usage*: A constructional approach to formal constraints on *go*-VERB. In Peter Uhrig and Thomas Herbst eds. *Yearbook of the German Cognitive Linguistics Association*. Berlin: Mouton de Gruyter, 231–252.

Hoffmann, Thomas. 2018. Review of Barðdal, Jóhanna, Elena Smirnova, Lotte Sommerer and Spike Gildea eds. 2015. *Diachronic Construction Grammar. Constructions and Frames* 10/1: 106–114.

Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Krug, Manfred. 2000. *Emerging English Modals. A Corpus-based Study of Grammaticalization.* Berlin: Mouton de Gruyter.

Lastres-López, Cristina. 2020a. Beyond conditionality: On the pragmaticalization of interpersonal *if*-constructions in English conversation. *Journal of Pragmatics* 157: 68–83.

Lastres-López, Cristina. 2020b. Subordination and insubordination in Contemporary Spoken English: *If*-clauses as a case in point. *English Today* 36/2: 48–52.

Light, Caitlin and Joel Wallenberg. 2015. The expression of impersonals in Middle English. *English Language and Linguistics* 19/2: 227–245.

Los, Bettelou. 2005. *The Rise of the* To-*Infinitive*. Oxford: Oxford University Press.

Loureiro-Porto, Lucía. 2009. *The Semantic Predecessors of* Need *in the History of English (c750–1710).* Oxford: Wiley-Blackwell.

Loureiro-Porto, Lucía. 2010. Verbo-nominal constructions of necessity with *þearf* n. and *need* n.: Competition and grammaticalization from OE to eModE. *English Language and Linguistics* 14/3: 373–397.

Lyons, Christopher. 1986. The syntax of English genitive constructions. *Journal of Linguistics* 22/1: 123–143.

Mair, Christian. 1997. Parallel corpora: A real-time approach to the study of language change in progress. In Magnus Ljung ed. *Corpus-Based Studies in English.* Amsterdam: Rodopi, 195–209.

Miura, Ayumi. 2015. *Middle English Verbs of Emotion and Impersonal Constructions: Verb Meaning and Syntax in Diachrony.* Oxford: Oxford University Press.

Möhlig-Falke, Ruth. 2012. *The Early English Impersonal Construction: An Analysis of Verbal and Constructional Meaning*. New York: Oxford University Press.

Payne, John and Eva Berlage. 2014. Genitive variation: The niche role of the oblique genitive. *English Language and Linguistics* 18/2: 331–360.

Payne, John and Rodney Huddleston. 2002. Nouns and noun phrases. In Rodney Huddleston and Geoffrey K. Pullum eds. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press, 323–523.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language.* London: Longman.

Reid, Dan. 2009. The great Festschrift makeover. *Addenda & Errata*. http://addenda-errata.ivpress.com/2009/03/the_great_festschrift_makeover.php (11 March, 2009.)

Rosenbach, Anette. 2002. *Genitive Variation in English: Conceptual Factors in Synchronic and Diachronic Studies.* Berlin: Mouton de Gruyter.

Ross, Daniel. 2018. Small corpora and low-frequency phenomena: *Try and* beyond contemporary, standard English. *Corpus* 18: 1–40.

Schneider, Edgar W. 2007. *Postcolonial English: Varieties around the World.* Cambridge: Cambridge University Press.

Sommerer, Lotte. 2015. The influence of constructions in grammaticalization. Revisiting category emergence and the development of the definite article in English. In Barðdal, Jóhanna, Elena Smirnova, Lotte Sommerer and Spike Gildea eds. *Diachronic Construction Grammar*. Amsterdam: John Benjamins, 107–137.

Sommerer, Lotte. 2018. *Article Emergence in Old English. A Constructionalist Perspective.* Berlin: Mouton de Gruyter.

Traugott, Elizabeth Closs and Graeme Trousdale. 2013. *Constructionalization and Constructional Change.* Oxford: Oxford University Press.

van der Gaaf, Willem. 1904. *The Transition from the Impersonal to the Personal Construction in Middle English*. Heidelberg: C. Winter.

van Gelderen, Elly. 2007. The definiteness cycle in Germanic. *Journal of Germanic Linguistics* 19/4: 275–308.

Van linden, An. 2015. Comparative modals: (Dis)similar diachronic tendencies. *Functions of Language* 22/2: 192–231.

Van de Velde, Freek. 2014. Degeneracy: The maintenance of constructional networks. In Ronny Boogaert, Timothy Colleman and Gijsbert Rutten eds. *Extending the Scope of Construction Grammar*. Berlin: Mouton de Gruyter, 141–179.

Warner, Anthony. 1982. *Complementation in Middle English and the Methodology of Historical Syntax. A Study of the Wyclifite Sermons*. London: Croom Helm.

Watanabe, Akira. 2009. A parametric shift in the D-system in Early Middle English: Relativization, articles, adjectival inflexion, and indeterminates. In Paola Crisma and Giuseppe Longobardi eds. *Historical Syntax and Linguistic Theory*. Oxford: Oxford University Press, 358–374.

Reviewed by
Teresa Fanego
Facultade de Filoloxía
Universidade de Santiago de Compostela
Avda. de Castelao s/n
E-15782 Santiago de Compostela
Spain
e-mail: teresa.fanego@usc.es