# RiCL

## Research in Corpus Linguistics

RiCL 7 (2019)

aelinco
Asociación Española de Lingüística de Corpus

# RiCL 7 (2019)

**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

# 7 (2019)





RiCL Research in Corpus Linguistics

# Foreword from the Editors

Paula Rodríguez-Puente - Carlos Prado-Alonso
University of Oviedo / Spain

It is our greatest pleasure to announce the publication of volume 7 of *Research in Corpus Linguistics* (RiCL), the official journal of the *Spanish Association for Corpus Linguistics* (AELINCO). In this volume we take over as new Editors-in-Chief of RiCL and we would not like to miss the chance to express our gratitude to all those who have made possible its publication.

First and foremost, we are grateful to the former editors, Javier Pérez-Guerra (University of Vigo) and María José López-Couso (University of Santiago de Compostela), who not only did an impeccable work launching and breeding RiCL during its early stages, but were also kind enough to guide us during our take-over and to provide valuable help at all times.

Thanks are also due to AELINCO, the sponsor of RiCL, for relying on us to maintain the high quality of the journal by fostering the publication of new and original contributions. We are specially thankful to Pascual Cantos (University of Murcia), Miguel Fuster (University of Valencia), Antonio Moreno (University of Málaga) and Marisa Carrió (Polytechnic University of Valencia), who conformed the Executive Board of the association at the time of our nomination. We also owe our deepest gratitude to all the AELINCO members who voted positively for our nomination during the last business meeting at the *XI International Conference of Corpus Linguistics* (CILC XI) held in Valencia on 16 May 2019.

Likewise, we feel greatly indebted to the anonymous reviewers for their timely and careful revisions, and to Zeltia Blanco-Suárez (University of Cantabria) who, as Assistant Editor of RiCL, provided valuable help and support.

Last but nonetheless least, we are grateful to the authors participating in volume 7, without whose contributions this volume would not have been possible. They have provided us with eight original articles on diverse topics related to the field of Corpus Linguistics, which no doubt will help us broaden our knowledge of this discipline.

We are absolutely thrilled about our first volume as Editors-in-Chief of RiCL and we would like to invite future authors, either members or non-members of AELINCO, to submit their contributions to the journal.

Oviedo 15 November 2019

RiCL Research in Corpus Linguistics

# How to build a corpus for a tool-based approach to determinologisation in the field of particle physics

Julie Humbert-Droz - Aurélie Picton - Anne Condamines
University of Geneva / Switzerland & University of Toulouse 2 / France
University of Geneva / Switzerland
CNRS & University of Toulouse 2 / France

**Abstract** – This paper discusses corpus design and building issues when dealing with a complex, multidimensional phenomenon such as determinologisation. Its representation in corpus data imposes an original reflection on the process and on some essential concepts of corpus building. This paper focuses on the necessity of representing the progressive aspects of determinologisation in the corpus, i.e. through levels of specialisation and through time, and the practical issues this raises. At the same time, it shows that a representative corpus of determinologisation in a specific domain (in this case, particle physics) implies clear and objective criteria when it comes to picking individual texts. Four principles are established to this end. The discussion leads to the proposal of a solid text selection procedure, which ensures that the peculiarities of determinologisation in the domain of particle physics are reflected in the corpus.

**Keywords** – corpus-building; determinologisation; comparable corpora; tool-based approach; representativeness; textual terminology

## 1. INTRODUCTION

The increasing use of corpora in linguistics and terminology has made it possible to address various research topics, such as terminological variation (e.g. Freixa 2006; Fernández-Silva 2016; Drouin *et al.* 2017), term and relation identification (e.g. Drouin 2003; León-Araúz *et al.* 2016; Daille 2017), circulation of terms outside of experts' sphere (e.g. Ungureanu 2006; Nicolae and Delavigne 2013; Condamines and Picton 2014). Considering the importance of corpora, numerous research papers and textbooks continually address typical issues related to corpus design (e.g. Biber 1993; Meyer and Mackintosh 1996; Kennedy 1998; Pearson 1998; Habert 2000; Ahmad and Rogers 2001; Bowker and Pearson 2002; McEnery and Hardie 2012). In this view, it is usually

argued that the design of a corpus needs to be thought out in accordance with the purpose for which the corpus is built. This means that the material included in the corpus must reflect the complexity of the phenomenon investigated.

In this context, this paper aims to discuss corpus design and building issues when dealing with a complex, multidimensional linguistic phenomenon such as determinologisation, studied from the viewpoint of one domain (particle physics), in French. We argue that the specificities of determinologisation impose an original and renewed reflection about corpus design, especially with regard to the issue of representativeness. The discussion leads to the building of one corpus, which will be referred to as *PPC* (*Particle Physics Corpus*).

This paper is structured as follows: Section 2 gives a brief overview of the background of the study for which the PPC is built. In Section 3, we attempt to operationalise the concept of representativeness and present the principles that were developed to this end. Section 4 outlines some concluding remarks and states the challenges that lie ahead for the exploration of the corpus.

## 2. BACKGROUND

### *2.1. Analysing determinologisation*

Determinologisation can be understood both as the process by which terms move from specialised language (LSP) into everyday language and as its result, i.e. the use of terms in a non-specialised context (Guilbert 1975; Meyer and Mackintosh 2000; Ungureanu 2006). In the latter case, it is known that semantic changes are likely to occur, such as the appearance of a shallower meaning, or metaphors, or word play (Meyer and Mackintosh 2000; Renouf 2017). As for the process, it can be considered as continuous on two levels. First, terms probably do not move into non-specialised language directly. Rather, they might be used in different genres and different levels of LSP communication in the process (Halskov 2005; Condamines and Picton 2014). Second, terms progressively integrate general language over time (Dury 2008; Renouf 2017).

In this context, our research aims to gain a broader understanding of determinologisation as a continuous process, one aspect that has received less attention

than the use of terms in non-specialised texts.[1] Our purpose is twofold: on the one hand, we aim to identify different factors that cause a term to determinologise, and on the other hand, we seek to better understand the role of the various media through which terms reach general language. In fact, the PPC is the first step of this research project. Its design results from a thorough analysis of the way in which determinologisation works, the diversity of texts involved in the process, and their representation in textual data.

### *2.2. A textual terminology methodology*

This approach is based on the principles of Textual Terminology (e.g. Bourigault and Slodzian 1999; Condamines 2003), in which the analysis is usually conducted on texts and in collaboration with domain experts. Such importance is given to textual data because "it is in the texts produced or used by a community of experts that most of the knowledge shared by this community is expressed, and thus accessible. Therefore, the analysis must begin there"[2] (Bourigault and Slodzian 1999: 30). From this viewpoint, specialised texts, usually gathered in corpora, constitute the primary material on which linguistic analyses are carried out, mostly from a tool-based approach. This approach mainly relies on comparable corpora, in which linguistic clues that are associated to the phenomena under study are explored, and the organisation of the data in sub-corpora is determined by the research purpose.

The differences that emerge from sub-corpora comparisons are interpreted in relation to the research purpose, and with the help of domain experts. In reality, since the analyst is usually not an expert of the domain under study, domain experts contribute to the analysis from the corpus compilation to the interpretation and validation of results. The whole analysis is thus a 'co-construction' process (Picton 2011: 137).

### *2.3. Particle physics as a 'determinologisable' domain*

In order to conduct a systematic analysis of determinologisation in a domain, one must first ensure that terms from this domain are likely to integrate general language.

[1] See Meyer and Mackintosh (2000), Ungureanu (2006), Renouf (2017) for such studies.
[2] Our translation.

Following Guilbert (1975: 84), we assume that such terms belong to domains that regularly appear in the media. Many domains could satisfy this condition, but we believe that particle physics is particularly relevant given the rather recent mediatisation of the building and exploitation of the LHC (Large Hadron Collider) as well as the Higgs boson discovery.

## 3. Representing determinologisation in corpus data: Four essential principles

In the previous section, we gave a brief overview of the study for which this corpus is built as well as the theoretical and methodological context. We will now explain how the concept of representativeness can be operationalised through a solid compilation method.

It must be pointed out, though, that because this concept has been extensively discussed in numerous research papers,[3] we will not further debate it. Rather, our point is that representativeness is an ideal that should be approached. To do so, we developed a strategic and informed decision-making process, which deals with the necessary heterogeneity of the data. Indeed, representing determinologisation in corpus data implies the inclusion of texts from different levels of specialisation, different genres and different time periods. In addition, our research project being restricted to a specific domain, the presence of relevant terms in the corpus must be ensured. This is achieved through a compilation method that relies on four principles. At this point, let us underline that, although this method is applied to French, it is language independent and can therefore be adapted to any other language. Some choices may differ, especially when it comes to identifying relevant genres, but the basic principles described in this paper remain valid.

### *3.1. From highly specialised language to general language*

The first principle relates to the level of specialisation of the texts. All the levels involved in the determinologisation process are to be considered, from highly specialised to general language, and the criteria established to determine these levels are discussed here.

[3] For example, Sinclair (1991), Biber (1993), Kennedy (1998), Habert (2000), Leech (2007).

First, it seems obvious that highly specialised language and general language should be included. The former is often represented by texts from what Bowker and Pearson (2002: 28) call an expert-expert level of LSP communication but the latter is a more complex concept (e.g. Ahmad and Rogers 2001: 735). Besides, because representing general language in a corpus is even more complex,[4] newspaper corpora are often considered to be an adequate operational choice (e.g. Halskov 2005; Dury 2008; Renouf 2017).

Second, let us consider the other two levels of LSP communication identified by Bowker and Pearson (2002: 28), which seem to be the most relevant to describe a communication level that is 'in between' (not highly specialised and not general). These are from experts to semi-experts and from experts to non-experts. The difference between semi-experts and non-experts mainly relies on people's knowledge of a subject. Semi-experts are considered to have some knowledge of a domain, whereas non-experts are considered to have none (or almost none). In reality, the difference might be subtler and more difficult to assert because it depends on the knowledge each individual has of a subject. Considering that the readership of a text might be very diverse, it seems even more difficult to determine the level of (non-)expertise of each individual. Therefore, although this distinction between semi- and non-experts is clear in theory, it does not seem fully operational here. This is why we would rather not distinguish between these two and include them both in the term 'intermediate level of specialisation'.

Texts from these three different levels of specialisation are essential to represent one progressive aspect of determinologisation and at least three sub-corpora should compose the corpus (one for each level identified). Let us now examine which genres are likely to best represent this process for each of these levels.

### *3.2. The importance of text genres*

According to Bhatia (2004: 23), "genre essentially refers to language use in a conventionalized communicative setting in order to give expression to a specific set of communicative goals of a disciplinary or social institution, which give rise to stable structural forms by imposing constraints on the use of lexico-grammatical as well as

[4] According to the large number of criteria extensively discussed for general language corpora (e.g. Sinclair 1991; Kennedy 1998; Siepmann *et al.* 2017).

discoursal resources." Many genres are found in LSP, which can be rather diverse according to both the level of specialisation and the domain (Meyer and Mackintosh 1996: 270). In our case, though, only the genres that are relevant for our research purposes must be selected, despite their diversity. Their identification relies on three main principles: 1) the genres must be likely to participate in the transfer of terms into general language, 2) they must be consistent with the levels of specialisation identified, and 3) they must be relevant for the domain under investigation.

In the following, and for explanatory purposes, we will first discuss the genres that compose the specialised and the non-specialised parts of the PPC and then those that are included in the intermediate part of the PPC.

First, according to Loffler-Laurian (1983: 10*sqq.*) and Bowker and Pearson (2002: 28), specialised articles are particularly relevant to represent highly specialised language. However, since there is a lack of this type of publication in French, doctoral theses were also considered.

Second, following the majority of authors who studied determinologisation from a corpus linguistics viewpoint, we compiled a corpus of general newspaper articles as a way to represent non-specialised language. In addition to the practical reasons discussed in 3.1, this choice implies a deeper conceptual motivation. On the one hand, Meyer and Mackintosh (2000: 112) argue that determinologisation describes the phenomenon that occurs "when a term captures the interest of the general public." It is assumed that this interest is reflected in the topics addressed by the media, hence in the terms they use. On the other hand, it is well known that the media are of great influence in this process (e.g. Cabré 1994: 593; Pearson 1998: 26; Moirand 2007: 20).

Third, the identification of genres that are relevant for the intermediate part of the corpus requires two additional conditions, which are complementary to the principles stated above. They are mostly based on two studies in which such texts are exploited: Condamines and Picton (2014: 171*sqq.*), who compiled a corpus of press releases, and Halskov (2005: 54*sqq.*), who used a corpus composed of science popularisation articles and 'newsgroup postings', among other genres. According to these studies, various genres from an intermediate level of specialisation are likely to play a part in the transfer of terms into general language, either directly or in a more indirect way. In this context, we believe that different genres should be included in order to best represent this diversity.

Therefore, the first condition is that anyone must be able to find and read the texts; they must not be restricted to a certain community. This will be referred to as the 'availability' condition. The second condition is based on the concept of 'knowledge transfer discourse'[5] (Beacco and Moirand 1995). According to the authors, many genres participate in transferring knowledge, even when it is not their primary purpose. Moreover, since knowledge is usually transferred through terms, terms probably appear in these genres, making them particularly relevant for our study. Thus, any genre that is described as a type of knowledge transfer discourse is considered as relevant for the intermediate part of the corpus. These conditions allow us to disregard genres such as textbooks, which seem to be restricted to a rather well delimited speech community. Genres such as press releases, general reports and science popularisation websites and articles appear to be much more adequate, as we explain below.

First of all, since press releases are by definition intended for journalists (Nicolae and Delavigne 2013: 219), journalists are likely to reuse the terms they find in press releases (Condamines and Picton 2014: 171*sqq.*). In this view, they represent a step in the transfer of terms from LSP to general language, more precisely from experts to journalists, and are thus relevant.

Secondly, we included general reports from several particle physics research laboratories. Annual (or biennial) general reports aim to inform the public about research activities. As such, they are considered as a type of knowledge transfer discourse. Moreover, since these reports are usually freely available online and may be read by anyone, the availability condition is also satisfied.

Lastly, we took into consideration two science popularisation genres. According to Guilbert (1975), for example, science popularisation and determinologisation are two closely related concepts, though the link between them is not clear. Authors such as Jacobi (1986) or Delavigne (2001) argue that science popularisation is an intermediary between experts and non-experts, with its main purpose being to transmit knowledge (Delavigne 2001: 28). As such, popularisation genres are particularly relevant. Moreover, to better represent the diversity of science popularisation media, we included two complementary genres: articles and websites. Indeed, journal articles are likely to treat current topics, such as news or discoveries, whereas websites tend to explain a domain in a more general way.

[5] Our translation.

To sum up, considering the arguments that were advanced in this section, the specialised part of the corpus is composed of specialised articles and theses, the intermediate part includes press releases, general reports, popularisation articles and websites, and the non-specialised part contains general newspaper articles.

### *3.3. Representing both dimensions of progression through a double division into sub-corpora*

The main challenge of this corpus design lies in finding a way to represent both progressive aspects of determinologisation, through levels of specialisation and through time. Given that our approach relies on a comparable corpus, this third principle is about the organisation of the data in sub-corpora. More precisely, we argue that two types of sub-corpora are necessary to reflect both dimensions of progression.

#### 3.3.1. First dimension of progression: Through levels of specialisation

As we mentioned in Section 3.1, the corpus should consist of at least three sub-corpora, one for each level of specialisation. However, in Section 3.2 we identified different genres that are relevant for these levels and for our research purposes. Therefore, based on the assumption that terms might behave in specific ways according to the genre, considering these genres separately seems more relevant. As a matter of fact, at this point, we do not know if the behaviours that we assume we will observe are related to the genre in which the terms are used, or to determinologisation – or both. Nevertheless, some genres should be grouped together, either because of their relative similarity and complementarity (popularisation articles and websites) or because of the practical reasons explained in 3.2 (specialised articles and theses).

| **Sub-corpus** | **Level of specialisation** | **Text genre** |
|---|---|---|
| Specialised | High | Specialised journal articles<br>Doctoral theses |
| Press releases (PR) | Intermediate | Press releases |
| Reports | Intermediate | Laboratory general reports |
| Science popularisation (SPop) | Intermediate | Journal articles<br>Websites |
| Press | Non-specialised | General newspaper articles |

Table 1: Composition of the PPC

As a result, the PPC is composed of five sub-corpora, which represent a way of approaching one continuous aspect of determinologisation, while remaining manageable in relevant corpus analysis tools[6] (see Table 1).

### 3.3.2. Second dimension of progression: Through time

The second dimension of progression is the diachronic dimension. For the data to reflect it, each of the five sub-corpora discussed above is further divided into smaller sub-corpora. The period to take into account and its division into shorter periods are discussed in this section.

According to Dury and Picton (2009: 38), when investigating evolution in recent or recently changing domains, it is probably more interesting to consider shorter periods, mainly because change can occur rather quickly. Picton (2011) calls such approach *short-term diachrony*. In order to determine these periods, two strategies are usually employed: the division is either arbitrary (e.g. several periods made up of the same number of years) or based on extra-linguistic criteria (Picton 2018: 44). In this case, we mainly rely on extra-linguistic criteria, which are related to the role of the media in determinologisation, and to the assumption that some important events of the domain might influence the ways in which terms are used in the corpus. Moreover, we assume that the media extensively covered these events, thus contributing to the transfer of terms in general language.

In accordance with the principles of Textual Terminology (Section 2.2), we collaborated with an expert to identify two events: the start of the LHC in 2008 and the discovery of the Higgs boson in 2012. Consequently, the corpus covers the period from 2003 to 2016 (2016 being the compilation time) and is organised in three shorter periods: 1) from 2003 to 2007, 2) from 2008 to 2011, and 3) from 2012 to 2016. Texts published prior to 2003 were not included so that the sub-corpora remain balanced and comparable. The corpus is thus composed of fifteen sub-corpora, as shown in Table 2.

[6] Indeed, it seems almost impossible to handle too many sub-corpora with current tools.

| | Progression through time | | |
|---|---|---|---|
| Progression through levels of specialisation | Specialised-2003-2007 | Specialised-2008-2011 | Specialised-2012-2016 |
| | PR-2003-2007 | PR-2008-2011 | PR-2012-2016 |
| | Reports-2003-2007 | Reports-2008-2011 | Reports-2012-2016 |
| | SPop-2003-2007 | SPop-2008-2011 | SPop-2012-2016 |
| | Press-2003-2007 | Press-2008-2011 | Press-2012-2016 |

Table 2: Composition of the corpus in terms of sub-corpora

### *3.4. Ensuring domain relevance through an objective text selection procedure*

This fourth principle addresses the issue of how each individual text is selected, which is a more operational viewpoint on corpus compilation. Indeed, not only should the texts be relevant for the determinologisation process (in terms of levels of specialisation, genres and publication dates), but they must be relevant for the domain as well. To this end, we detail a solid text selection procedure for the sub-corpora to remain balanced in terms of content. Paradoxically, this procedure must be flexible enough so that it can be adapted to the necessary heterogeneity of the documents.

For explanatory purposes, we first discuss the sub-corpora containing texts from either a high or intermediate level of specialisation. Second, we argue that this procedure must be refined for the *Press* sub-corpus.

#### 3.4.1. A balance between keywords and experts

One common method of assessing the relevance of texts when building a corpus is based on a usually quick evaluation of their content. According to Pearson (1998: 54), this may be achieved "by looking at what a particular text is about (e.g. on the basis of its title, table of contents in the case of a book)" and by "examining the lexical structure of a text and identifying keywords used frequently in the text." To this end, we developed an approach based on specific terms considered as key to the domain. Collaborating with an expert was necessary to define a sufficient number of keywords (almost) unequivocally referring to the domain, and in particular to the subdomain of

the Standard Model of particle physics.[7] She pointed at terms such as *Modèle Standard*, *boson de Higgs*, *ATLAS*, *LHC* or *particule élémentaire*,[8] and the texts containing them in titles, tables of contents or even sometimes in the body were retained.

In fact, the expert played a determining role throughout this corpus building process. Not only did she identify the relevant events used for the diachronic division, but she also clarified the complex links between CERN (European Organization for Nuclear Research), the Standard Model, the LHC and the Higgs boson, leading us to find the most relevant sources. Therefore, given that the aforementioned events both happened at CERN and that CERN is located at the Swiss-French border, only French and Swiss sources related to this organisation were included. Thus, all the texts come from:

- Swiss and French universities that provide access to theses in French,
- the only French research journal that publishes articles in French,
- Swiss and French websites of laboratories undertaking research in particle physics (including CERN),
- French science popularisation journals, and
- Swiss and French newspapers.

Based on these sources, the overall text selection procedure broadly consisted of 1) listing the individual texts containing at least one of the keywords, 2) discussing and refining the list with the expert, and 3) balancing the size of the diachronic sub-corpora so that they remained comparable. In other words, domain relevance is ensured both by the presence of certain keywords in the texts and by a close collaboration with a domain expert. However, this selection procedure is not adequate for the *Press* sub-corpus and it must be adapted, as we explain in the next section.

### 3.4.2. Refining the procedure for the particular case of the *Press* sub-corpus

Although our corpus design includes newspaper articles containing particle physics terms, our research purposes require more varied material. Since determinologised terms can behave in various ways in a non-specialised context, the possibility of finding such behaviours in the corpus must be ensured. To do so, however, the text selection

[7] Narrowing down the domain proved necessary given the large number of subdomains comprised in particle physics.
[8] *Standard Model, Higgs boson, ATLAS, LHC, elementary particle.*

procedure cannot rely on the same limited number of keywords as the other four sub-corpora. Otherwise, the articles would be very similar in content. More importantly, the different behaviours resulting from determinologisation would likely be missing.

That being said, a random selection does not seem operational either. Indeed, for particle physics terms to be analysed, we need to make sure that they appear in the corpus – and that they appear frequently enough. Thus, we propose a hybrid method that guarantees the presence of relevant terms, while ensuring that some occurrences are examples of determinologised terms. Articles are selected based on a large number of terms attested in the other four sub-corpora, which are used as keywords on the platform LexisNexis®.[9] Such a large number of keywords provides more diverse results and avoids the bias of selecting rather similar articles. Furthermore, a large number of keywords also maximises the chances of observing a whole variety of contexts, some of which might be linked to determinologisation. But the selection process must be carried out carefully, by choosing not only articles containing many keywords, but also, and more importantly, articles containing few. Indeed, if only one term appears in an article, for example, this one occurrence could be a metaphor or word play, or even another consequence of determinologisation that we do not know of yet.

In this context, an objective method was developed in order to identify the terms that are attested in the other sub-corpora and that are relevant to retrieve the articles on LexisNexis®. It is illustrated in Figure 1.

[9] It is accessed via a subscription at the University of Geneva.

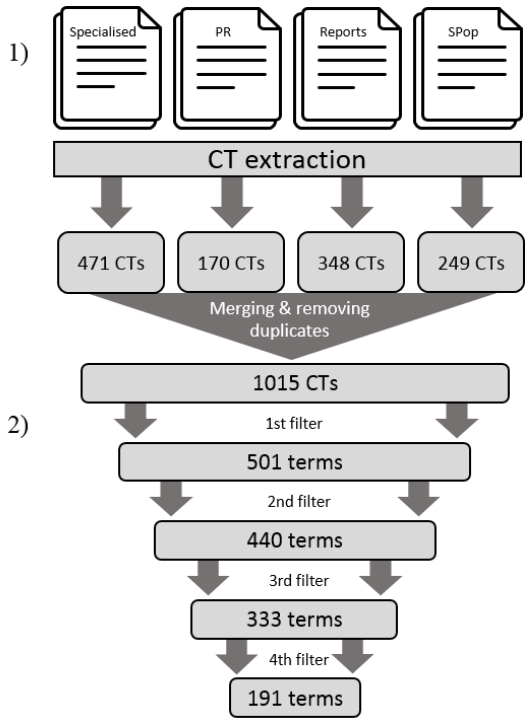


Figure 1: Term extraction and keyword selection

The method broadly consists of 1) a candidate term (CT) extraction and 2) a refinement of the list of CTs through objective filters. The CTs were extracted from the *Specialised*, *PR*, *Reports* and *SPop* sub-corpora with TermoStat (Drouin 2003) and only those with a specificity score higher or equal to 40 were retained. After removing the duplicates, the list consists of 1015 CTs.

As TermoStat is a hybrid term extraction system (Drouin 2003: 99), some of the extracted CTs were in fact noise, whereas other CTs seemed less relevant to build the *Press* sub-corpus. Thus, we applied four filters to keep the most relevant terms only and we removed the following:

a. **CTs that are actually not French terms**. They fall into three categories:

- proper nouns (e.g. *Perrine*, *Rolf*), countries (e.g. *République Slovaque*), abbreviations appearing in bibliographical references (e.g. *phys*, *nucl*), parts of URLs or email addresses (e.g. *lhc-france*, *grey@cern*);
- incomplete CTs, probably due to tagging inaccuracy (e.g. *détecteur de pied*, instead of *détecteur de pied de gerbe*); and

- English terms that are only used in the English parts of the texts, such as references (e.g. *calorimeter*, *polarization*).

b. **CTs that do not designate domain-specific concepts and that are considered as transdisciplinary** (such as *analyse*, *interaction*, *fonctionnement*). This filter is based on the *transdisciplinary scientific lexicon* (e.g. Tutin 2007; Drouin 2007), and we exploited the list established by the Scientext project.[10]

c. **CTs that belong to more than five domains in more than three terminological data banks** (Grand dictionnaire terminologique, Termium, FranceTerme, IATE and TERMDAT), such as *accélération*, *grille*, *collision*, *vitesse*. This step is based on the idea that, although these terms do belong to particle physics, they are likely to be polysemous in newspaper articles. As a result, some of their occurrences may neither designate a concept of particle physics nor convey a determinologised meaning of a particle physics term. Thus, they appear to be inadequate for the selection procedure discussed here.

d. **CTs that are not attested in all of the four sub-corpora**, such as *superchamp, préon, leptoquark.*

The final list is composed of 191 terms. We believe that this procedure allowed us to build a sub-corpus that is relevant for the domain – given that the articles contain one or more of these terms – and that is adequate to explore the consequences of determinologisation – given that they were anticipated through a balanced text selection. This last sub-corpus was the final step of the building process discussed in this paper and it completes the whole corpus (see Table 3).

| Sub-corpus | 2003-2007 | 2008-2011 | 2012-2016 | Total |
|---|---|---|---|---|
| Specialised | 314,658 | 330,975 | 349,242 | **994,875** |
| PR | 70,950 | 69,478 | 69,892 | **210,320** |
| Reports | 516,820 | 302,552 | 322,501 | **1,141,873** |
| SPop | 216,969 | 194,675 | 208,401 | **620,045** |
| Press | 367,378 | 365,650 | 365,680 | **1,098,708** |
| Total | 1,486,775 | 1,263,330 | 1,315,716 | **4,065,821** |

Table 3: Size of the corpus in number of occurrences

[10] Scientext project, https://scientext.hypotheses.org/, last access: 28 April 2019.

## 4. CONCLUDING REMARKS

In this paper, we presented an original reflection on the design of a corpus meant to be representative of the determinologisation process in particle physics. In particular, we discussed some essential issues regarding corpus building and the specific ways they must be addressed given the progressive dimensions involved in this process. From this viewpoint, we mainly discussed how the concept of 'representativeness' can be operationalised through an objective compilation method that relies on four principles:

1. the texts included in the corpus represent the levels of specialisation involved in the determinologisation process (highly specialised, intermediate, non-specialised);
2. they belong to genres that are likely to take part in this process. This feature was mainly identified based on the concepts of 'availability' and 'knowledge transfer';
3. the progressive aspects of determinologisation are represented by two types of sub-corpora. A division into five sub-corpora reflects progression through levels of specialisation, and each of these sub-corpora is divided into three diachronic sub-corpora, which represent progression through time;
4. the texts are relevant given the domain investigated. A solid and objective text selection procedure was developed to this end.

Our work now focuses on the proper exploration of the corpus. Indeed, such a large number of sub-corpora remains a challenge for any corpus analysis tool and for the analyst. If it is possible to take into account both dimensions involved in determinologisation separately, analysing them simultaneously, as well as their interactions, seems much more challenging. Finding methods that enable us to handle so many sub-corpora is therefore crucial in order to better understand determinologisation.

## REFERENCES

Ahmad, Khurshid and Margaret Rogers. 2001. Corpus linguistics and terminology extraction. In Sue E. Wright and Gerhard Budin eds. *Handbook of Terminology Management*. Amsterdam: John Benjamins, 725–760.

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

Beacco, Jean-Claude and Sophie Moirand. 1995. Autour des discours de transmission des connaissances. *Langages* 117: 32–53.

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

Bourigault, Didier and Monique Slodzian. 1999. Pour une terminologie textuelle. *Terminologies Nouvelles* 19: 19–32.

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

Cabré, M. Teresa. 1994. Terminologie et dictionnaires. *META* 39/4: 589–597.

Condamines, Anne. 2003. *Sémantique et Corpus Spécialisés: Constitution de Bases de Connaissances Terminologiques*. Toulouse: Université Toulouse le Mirail.

Condamines, Anne and Aurélie Picton. 2014. Des communiqués de presse du Cnes à la presse généraliste. Vers un observatoire de la diffusion des termes. In Pascaline Dury, José Carlos de Hoyos, Julie Makri-Morel, François Maniez, Vincent Renner and María Belén Villar Diaz eds. *La Néologie en Langue de Spécialité: Détection, Implantation et Circulation des Nouveaux Termes*. Lyon: Centre de Recherche en Terminologie et Traduction, Université Lumière Lyon 2, 165–188.

Daille, Béatrice. 2017. *Term Variation in Specialised Corpora*. Amsterdam: John Benjamins.

Delavigne, Valérie. 2001. *Les Mots du Nucléaire. Contribution Socioterminologique à une Analyse des Discours de Vulgarisation*. Université de Rouen dissertation.

Drouin, Patrick. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology* 9/1: 99–117.

Drouin, Patrick. 2007. Identification automatique du lexique scientifique transdisciplinaire. *Revue Française de Linguistique Appliquée* 12/2: 45–64.

Drouin, Patrick, Aline Francoeur, John Humbley and Aurélie Picton eds. 2017. *Multiple Perspectives on Terminological Variation*. Amsterdam: John Benjamins.

Dury, Pascaline. 2008. The rise of carbon neutral and compensation carbone: A diachronic investigation into the migration of vocabulary from the language of ecology to newspaper language and vice versa. *Terminology* 14/2: 230–248.

Dury, Pascaline and Aurélie Picton. 2009. Terminologie et diachronie: Vers une réconciliation théorique et méthodologique? *Revue Française de Linguistique Appliquée* 14/2: 31–41.

Fernández-Silva, Sabela. 2016. The cognitive and rhetorical role of term variation and its contribution to knowledge construction in research articles. *Terminology* 22/1: 52–79.

Freixa, Judit. 2006. Causes of denominative variation in terminology. A typology proposal. *Terminology* 12/1: 51–77.

Guilbert, Louis. 1975. *La Créativité Lexicale*. Paris: Larousse.

Habert, Benoît. 2000. Des corpus représentatifs: De quoi, pour quoi, comment? In Mireille Bilger ed. *Linguistique sur Corpus: Études et Réflexions*. Perpignan: Les Presses de l'Université de Perpignan, 11–58.

Halskov, Jakob. 2005. Probing the properties of determinologization: The DiaSketch. *Lambda* 29: 39–63.

Jacobi, Daniel. 1986. *Diffusion et Vulgarisation: Itinéraires du Texte Scientifique*. Paris: Les Belles Lettres.

Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics*. London: Longman.

Leech, Geoffrey. 2007. New resources, or just better old ones? The Holy Grail of representativeness. In Marianne Hundt, Nadja Nesselhauf and Carolin Biewer eds. *Corpus Linguistics and the Web*. Amsterdam: Rodopi, 133–149.

León-Araúz, Pilar, Antonio San Martín and Pamela Faber. 2016. Pattern-based word sketches for the extraction of semantic relations. In Patrick Drouin, Natalia

Grabar, Thierry Hamon, Kyo Kageura and Koichi Takenchi eds. *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*. Osaka, Japan, 73–82.

Loffler-Laurian, Anne-Marie. 1983. Typologie des discours scientifiques: Deux approches. *Études de Linguistique Appliquée* 51: 8–20.

McEnery, Tony and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Meyer, Ingrid and Kristen Mackintosh. 1996. The corpus from a terminographer's viewpoint. *International Journal of Corpus Linguistics* 1/2: 257–285.

Meyer, Ingrid and Kristen Mackintosh. 2000. When terms move into our everyday lives: An overview of de-terminologization. *Terminology* 6/1: 111–138.

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

Nicolae, Cristina and Valérie Delavigne. 2013. In Geoffrey Williams ed. *Actes des Sixièmes Journées de la Linguistique de Corpus*. Lorient: Université de Bretagne-Sud, 217–229.

Pearson, Jennifer. 1998. *Terms in Context*. Amsterdam: John Benjamins.

Picton, Aurélie. 2011. Picturing short-period diachronic phenomena in specialised corpora. A textual terminology description of the dynamics of knowledge in space technologies. *Terminology* 17/1: 134–156.

Picton, Aurélie. 2018. Terminologie outillée et diachronie: Éléments de réflexion autour d'une réconciliation. *ASp* 74: 27–52.

Renouf, Antoinette. 2017. Some corpus-based observations on determinologisation. *Neologica* 11: 21–48.

Siepmann, Dirk, Christoph Bürgel and Sascha Diwersy. 2017. The *Corpus de Référence du Français Contemporain* (CRFC) as the first genre-diverse mega-corpus of French. *International Journal of Lexicography* 30/1: 63–84.

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

Tutin, Agnès. 2007. Autour du lexique et de la phraséologie des écrits scientifiques. *Revue Française de Linguistique Appliquée* 12/2: 5–14.

Ungureanu, Ludmila. 2006. *L'Interpénétration Langue Générale-Langue Spécialisée dans le Discours d'Internet*. Paris: Connaissances et Savoirs.

*Corresponding author*
Julie Humbert-Droz
UNI MAIL
40, Bd du Pont-d'Arve
1211 Genève 4
Switzerland
e-mail: julie.humbert-droz@unige.ch

received: May 2019
accepted: September 2019

RiCL Research in Corpus Linguistics

# Vocabulary learning through data-driven learning in the context of Spanish as a foreign language

Gang Yao
University of Murcia / Spain

**Abstract** – An increasing number of studies have shown the potential associations between corpus work and second language acquisition and teaching. Some research, for example, explores the effect of data-driven learning (DDL, Johns 1991) in the context of foreign language learning. Up till now, however, empirical quantitative studies on the topic have been limited, especially with respect to foreign languages other than English. In order to bridge this gap, a quasi-experimental longitudinal design was used in the present study to examine whether there is a statistically significant difference between the DDL approach to vocabulary learning and more traditional learning methods (e.g., dictionary approach) in the context of Spanish as a foreign language (SFL) by Chinese students. The study further gauged students' attitude towards DDL activities. The results of two post-tests revealed that the DDL group of students significantly outperformed the group of students following a traditional learning method. Furthermore, a questionnaire assessment collected from the experimental group showed that the respondents generally favored DDL and adopted a positive attitude towards its future application to Spanish learning.

**Keywords** – Data-driven learning; Spanish as a foreign language; vocabulary learning; empirical study

## 1. INTRODUCTION[1]

Vocabulary is considered to be one of the most important elements when learning a foreign language (Nation 2001). Lewis (1993: 89) comments that "lexis is the core or heart of language […]." Nation (2001) and Nation and Meara (2010) state that knowledge of vocabulary enables language use, which can be reflected in all language skills (i.e., listening, speaking, reading, and writing). Similarly, both Schmitt (2000) and

[1] This research was presented at the *X Congreso Internacional de Lingüística de Corpus* (CILC2018) and is partly supported by a China Scholarship Council Grant within the Graduate Student Overseas Study program. The author would like to thank an anonymous reviewer and John Higgins for constructive and valuable comments.

Jiménez-Calderón and Sánchez-Rufat (2017) agree that lexical knowledge is fundamental to communicative competence and second language acquisition. Barcroft (2005) highlights the importance of vocabulary in three aspects: communication, perception, and the way in which the knowledge of grammar is stored in the mind.

Although we are conscious of its importance, there exist many challenges and problems in second language vocabulary acquisition, a fact that is especially true in Chinese SFL context. Vocabulary teaching and learning procedures are relatively simple and old-fashioned in China (Guan 2013). For instance, the example sentences used for vocabulary teaching are often extracted from traditional monolingual or bilingual dictionaries or, in some cases, teachers themselves invent or compile those examples. Such sentences, which may have little authenticity and contextual adequacy, are unlikely to arouse learner's interest and attention. Meanwhile, students still draw heavily on teacher's explaining and rote learning (Chang 2001). This top-down learning process weakens their initiative and autonomous learning since they tend to receive linguistic input passively.

With the advent and development of computer technology, computer-assisted language learning (CALL) has become possible and is maturing. One thing that affects CALL significantly is the use of digitized corpora. From the very first Brown Corpus to large-scaled modern corpora (e.g., COCA, WebCorp), corpora have been exerting a considerable influence on language teaching and learning in many ways (O'Keeffe *et al.* 2007, Szudarski 2018). As freely available and easy-to-access corpus resources come into being, the above-mentioned problems and challenges in second language vocabulary can be addressed from a fresh angle, namely, corpus-aided vocabulary learning (Guan 2013; Yılmaz and Soruç 2015; Karras 2016). And one of the most significant representations of corpus-aided learning is data-driven learning (DDL), proposed by Johns (1991).

## 2. Data-driven learning: Theoretical background

DDL is a new approach to language learning and teaching in which students can inductively discover linguistic features and regularities by exploring "real and authentic language data" (Johns 1991). Essentially, DDL exploits the techniques of corpus use in contrast to traditional learning strategies, in which textbook learning and teacher's

explanations feature prominently. But the problem is that users cannot read a corpus directly; instead, they need to draw upon a program or software, namely, a concordancer. Using this interface, one can retrieve concordance lines after requesting a word, a phrase or a regular expression in the search bar. The concordance lines are typically presented in the form of Key Word in Context (known as KWIC), with the keyword displayed centrally and some words before and after (cf. Figure 1). Observing the concordance formatted as KWIC, the learner can, for instance, easily discover patterns, be sensitive to collocations, and enhance their learning strategy (Thurstun and Candlin 1998; Pérez-Paredes 2010). Take the example in Figure 1, i.e., the Spanish verb *infringir*. With the co-text around the search word, it would not be difficult for a learner to notice that the word usually takes nouns related to the law as objects. Now we can glimpse the main characteristic of DDL, namely that learners themselves discover linguistic regularities and make generalizations about linguistic phenomena based on observation, analysis, induction, and conclusion. This type of learning is also called *discovery learning* (Bernardini 2000, 2004). After an amount of training, *the learner as researcher* (Johns 1991; McEnery and Wilson 1997; Gavioli 2001) can engage in his own linguistic analyses and exploration.

| | | |
|---|---|---|
| Conocer si una determinada invención puede ponerse en práctica sin | **infringir** | los derechos derivados de una patente. |
| irracional, irrazonable, caprichoso, y que sea ilegal, esto es que | **infrinja** | abiertamente la legislación vigente. El artículo pertinente de la |
| En caso de que el incumplimiento | **infrinja** | alguna de las prohibiciones establecidas por la Ley del Sector Eléctrico |
| ¿Qué pasa en este país para que se mate, se viole, se robe y se | **infringe** | las leyes con el salvajismo que estamos padeciendo? ¿Qué ocurre en |
| fue maltratado por individuos de civil. La CNI lo acusó de | **infringir** | la Ley de Seguridad Interior del Estado, |
| la caballería andante se enfrentó con el poder del Estado. Don Quijote | **infringió** | sin que nadie le pidiera cuentas, leyes muy sensibles que prevén |
| La demanda por daños, presentada ayer, acusa a la Academia de | **infringir** | la ley de derecho de autor del personaje Blanca Nieves |
| Eres astuto, europeo. Sabes muy bien que no puedo | **infringir** | ese principio. Esgrimiéndolo, me entregas a ti atado de pies y manos. |
| "Aquel que viola la Constitución, aquel que | **infringe** | , aquel que clausura el Parlamento, aquel que clausura la democracia, |
| el truchimán de Morales estuvo a punto de mojarme la oreja, de | **infringir** | me una humillante derrota en una apasionada discusión sobre la radio |
| La Fox | **infringe** | esas normas, ya que posee en la actualidad el 99% de las acciones |
| reconocemos derechos a ciertas criaturas, aceptamos la obligación de no | **infringir** | los en nuestro trato con ellas. |
| El Código de circulación prevé fuertes multas para los que | **infringen** | sus normas. |
| Quienes | **infrinjan** | a prohibición establecida en este Acuerdo, serán juzgados |
| el electrón no sólo huye y desaparece, sino que al hacerlo | **infringe** | todas las leyes conocidas. |
| es decir, irracional, irrazonable, caprichoso, y que sea ilegal, esto es, que | **infrinja** | abiertamente la legislación vigente. |
| Conocer si una determinada invención puede ponerse en práctica sin | **infringir** | los derechos derivados de una patente. |
| anunció la presentación de una querella contra la compañía española por | **infringir** | derechos constitucionales». |
| en los años setenta es una política cultural en la que todo lo que | **infrinja** | la "coherencia narrativa" puede ser interpretado como radical. |
| proseguir el combate. En el séptimo asalto Hurtado le | **infringió** | otra herida en la frente por un golpe "no intencional" y en cambio |

Figure 1: Concordance of 'infringir' retrieved from the 'Corpus de Referencia del Español Actual' (CREA) by the Royal Spanish Academy

However, there arises another problem: How to train our learners to become familiar with DDL? Teachers play a role here: not a dominant one but as "director and coordinator" (Johns 1991: 3), designing concordance-based exercises to provide their students with practice. According to Boulton (2010a, 2010c), there are two types of

DDL exercises: one is hands-on DDL, which corresponds to "direct corpus consultation" (Chambers 2007: 4); the other one is hands-off DDL, which is equivalent to "indirect corpus consultation." In the first type of DDL, students are given considerable individual autonomy, which can be considered as a 'pure' DDL. However, in the case of inexpert learners, for instance, the extensive data retrieved from corpora may frustrate them because the data sometimes are "irrelevant", "incomprehensible", and "extremely chaotic" (Boulton 2010c: 6). Besides, corpus data are normally produced by native speakers, and thus the difficulty may lie outside the language competence of a learner. Lastly, direct corpus consultation requires access to a computer, which means learners have to know basic computer skills and schools may need to be equipped with multimedia rooms. But this condition is hard to achieve for some students and education centers (Pérez-Paredes 2005). Conversely, the hands-off DDL, a "soft version" (Gabrielatos 2005), is popularly introduced into regular classrooms as corpus resources. It is the teacher who consults the corpora directly, then selects appropriate language data depending on the learner's level, and finally prepares concordance-based exercises (e.g., handouts, worksheets). The advantages of this type of DDL may be easily noticed: students can still stand to benefit from direct access to authentic language data (instead of direct corpus consultation). At the same time, computer knowledge and competencies are not required. Thompson (2006) also points out that selected materials can help our learners turn their attention to the key elements, reduce confusion, and confine the range of possible answers (cf. also Stevens 1991). For newcomers to the corpus, preselected concordance lines make linguistic features more noticeable (cf. Sripicharn 2010). Moreover, prepared materials could reduce cognitive load at the beginning since learners only need to focus on a single new element (Boulton 2010a).

## 3. Previous studies on DDL

The use and effects of DDL for language learning and teaching have been studied thoroughly and systematically by many researchers. The emergence of several holistic surveys and syntheses suffices to show the trends and popularity of DDL, such as Chambers (2007), Boulton (2008, 2010b, 2017a), Boulton and Cobb (2017), Mizumoto and Chujo (2015), and Lee *et al.* (2019). All of them have given a fair summary of previous studies on DDL, and some of them have even conducted a meta-analysis. The

state-of-the-art review in this section will only focus on empirical studies which, in turn, can be divided into: qualitative analyses and quantitative analyses. It should be noted that there does not exist a strict boundary between these two categories. In practice, many studies adopted both methods to complement one another, i.e., mixed methods. Labeling a study qualitative only means the study has more qualitative characteristics than quantitative ones; and vice versa.

### *3.1. Qualitative studies*

As stated in Boulton (2017b: 185), many initial publications related to DDL lie in "emic studies", with the goal of exploring what learners think about DDL; in other words, the evaluation of DDL. Specifically, there are three types of evaluation (Boulton 2008, as cited in Gilquin and Granger 2010: 365): evaluation of attitudes (what do learners think about DDL?), practices (how well do users work with DDL?), and efficiency (can learners really benefit from DDL?). In emic studies, information of this kind is usually collected through interviews, learning logs, and especially questionnaires.

The focus of this type of empirical study is usually on learners' written production with the aim of improving writing skills (e.g., Chambers and O'Sullivan 2004; O'Sullivan and Chambers 2006; Kennedy and Miceli 2010; Charles 2012; Chang 2014). Yoon and Hirvela (2004), for example, examined corpus use in students' L2 writing and their perceptions of it. The authors combined qualitative and quantitative analysis in their study. The feedback from the students was generally positive and most learners indeed favored corpus-assisted writing. In particular, corpus assistance was deemed beneficial in terms of learning common usage and collocates and boosting student's confidence in writing. There were, however, several problems or difficulties reported in the study, such as the time that was wasted on the corpus searches and students' proficiency level as an essential factor in corpus work.

Qualitative studies on vocabulary acquisition are relatively new. Jiao (2012) utilized corpora to help students ($N$ = 87) learn English vocabulary. After a one-semester instruction with the aid of corpora, all participants were invited to take a survey. Again, the students generally acknowledged the merit of corpus work since it can contribute to autonomous learning and help the students grasp the correct usage of vocabulary (collocation, colligation, semantic prosody, etc.). Likewise, Tekin and Soruç

(2016) enabled 26 participants from a Turkish high school to use BNC to learn four target words and then received their reflective journals. The qualitative findings showed that the students considered corpus-assisted vocabulary learning activities interesting, innovative, practical but also complex. Aşık *et al.* (2016) reported 126 Turkish EFL learners' perceptions of DDL regarding lexical awareness and development. Although the data they collected was based on questionnaires and interviews, the authors quantified it to carry out a statistical analysis. The results revealed that the students held overall positive opinions about DDL tasks, although improvements can only be seen in certain aspects of lexical awareness, such as synonyms and collocations; while awareness concerning word frequency, idioms, and learning strategies did not achieve a satisfactory result.

Despite the fact that the majority of studies above lack quantitative data and statistical analyses, it does not mean they are of little value. On the contrary, during the initial phase of the development of DDL, qualitative research on learners' attitudes and behaviors is undoubtedly helpful for other researchers who want to know what has been done so far in this field and what are the advantages/problems of DDL (Chambers 2007).

### *3.2. Quantitative studies*

Even though DDL has been developing and perfecting since the 1990s, we should address ourselves to some key questions: Does DDL indeed work for foreign language learning? To what extent it is effective? To answer these questions, it is necessary to provide more empirical evidence that focuses on measurable outcomes in order to shed light on the effectiveness and efficiency of DDL. That is, quantitative studies that observe DDL from a "more etic perspective" (Boulton 2017b: 186) are needed.

Within quantitative studies, according to Boulton (2017a, 2017b), two categories can be identified depending on the purpose of corpus use. But, again, there is no watertight delimitation between them. The first group aims to evaluate the effect of the corpus as a reference resource, particularly corpus use in practice exercise, translation or learner's written revision. Interestingly, most studies that fit into this category are qualitative (cf. Section 3.1). The reason behind this may be that it is difficult to quantify

the results of writing or translation evaluation. Notwithstanding, there are several quantitative studies that employ corpus as a reference resource.

For example, Gaskell and Cobb (2004) carried out a longitudinal experiment in an intermediate-low level English writing course. A total of 20 learners of English were involved in this experiment for over 15 weeks. The task consisted mainly of error identification, corpus consultation with instructor's aids, and independent searches in the corpus. The comparison of students' writings between the pre-test and post-test, as well as a questionnaire assessment, indicated that all students thought they had achieved improvements in grammar and error correction after the course. In Gilmore's (2009) short-term study, 45 intermediate-level Japanese learners of English first received a 90-minute training session to solve lexical and grammatical problems they had encountered in their writings. Their second compositions, which were graded by four native speakers of English, showed a significant improvement in terms of naturalness. Crosthwaite (2017) examined the adequacy of corpus use for student error correction in L2 writing during a series of DDL course. Teachers offered error feedback while students highlighted revisions made with corpus consultation or without it. The quantitative results revealed that with corpus-mediated correction, students can prevent lexical errors more successfully, but they were less likely to correct morphosyntactic errors. Students' feedback from the post-course questionnaires also confirmed the quantitative results. A similar writing enhancement experiment, conducted by Cotos *et al.* (2017), incorporated a corpus-based platform—Research Writing Tutor (RWT)—into a one-semester writing course. The RWT can automatically evaluate students' drafts and give them rhetorical feedback. Multiple comparisons in a mixed-methods design revealed that RWT-enabled DDL activities could improve the quality of students' writing, for instance in genre awareness.

The other group of studies that focus on the effect of the corpus as a learning aid shows an interest in examining how the corpus can assist learners in linguistic elements of language learning such as vocabulary and grammar. Stevens' (1991) study could be considered a pioneering work in this sense. He innovatively used concordance-based exercises, instead of conventional gap-fillers, to aid students in vocabulary learning. The result suggested that this new type of exercises was easier and more useful for learners and that it can become a viable alternative to the traditional exercises. Cobb (1997, 1999) put DDL into practice in a stricter sense by designing and introducing a

well-known web-based platform—Compleat Lexical Tutor.[2] In his experiments, one group of students were asked to learn 240 English words based on that interactive platform during one semester, while the other group of students followed a traditional teaching method with a dictionary. The findings suggested that both treatments were effective for the acquisition of word meaning in a short period, but only the experimental group performed significantly well on the retention of vocabulary for an extended period.

Entering the 21st century, studies on vocabulary learning through DDL start to appear. For example, Allan (2006) carried out an experiment in which 18 advanced learners of English were engaged. The experimental group ($N$ = 13) was given the concordance-based task to learn vocabulary for over 12 weeks. A quantitative analysis of the results indicated that the DDL group outperformed the other group, although the conditions of the two groups during the experiment were not entirely comparable. A more in-depth study was carried out by Anani-Sarab and Kardoust (2014), who investigated the potential and implication of corpus in the context of English as a Foreign Language (EFL) in an experiment with 34 Iranian students who were preparing for an English test. The experimental group adopted DDL activities to learn phrasal verbs, while the control group did the same through dictionary-based activities. After 14 sessions of instruction, the results from the immediate and delayed post-tests showed that the DDL group achieved greater improvements. However, this study did not include students' assessment of DDL activities in the experiment, which is popularly considered an important aspect in research of this kind.

Yılmaz and Soruç (2015) and Soruç and Tekin (2017) examined the effectiveness of DDL on vocabulary learning and teaching by contrasting the concordance-based vocabulary instruction with the traditional instructions (such as dictionary definitions, synonyms, fill-in-the-blank exercises). Despite the difference in settings, both experiments reported that DDL vocabulary learning activity yielded better results after comparing the pre-test and the post-test. Soruç and Tekin (2017) integrated a further delayed post-test to compare it with the immediate post-test. The result also supported the superiority of DDL. Interviews from both experiments also recorded students' positive attitudes towards it. Nonetheless, the procedure of Yılmaz and Soruç's (2015) study was not entirely clear. For example, the duration of the DDL activity and its retention effect

[2] Compleat Lexical Tutor: https://www.lextutor.ca/.

were unknown. Differing from other research on vocabulary learning through DDL, Karras (2016) conducted a more large-scale study, in which 100 international students of a Vietnam secondary school participated over eight weeks. In this study, both the experimental and control groups had online dictionary learning activities, but the first group received an extra DDL training. Based on the weekly results, the author reported that both groups achieved improvements but the DDL group obtained significantly higher scores than the other group. It is worth noting that the factor of grade level influenced the effect of DDL. However, since the two groups received a different amount of treatment, it cannot be said that they were comparable.

In addition to vocabulary learning, corpus work has been proved helpful in learning other linguistic features. For instance, corpus-driven lexico-grammatical learning has been shown to provide favorable outcomes, primarily on collocation and colligation (Chan and Liou 2005; Koosha and Jafarpour 2006; Huang 2014; Daskalovaska 2015; Vyatkina 2016; Li 2017). Boulton (2009), Smart (2014), and Moon and Oh (2017), in turn, paid more attention to aspects of grammar such as the passive voice, the overuse of *be*. All these studies indicate that DDL activities significantly improve learners' grammatical capacity.

More recently, the application of DDL is diversifying; in other words, DDL research does not only cover linguistic features (e.g., vocabulary, grammar, writing), but also shifts attention towards language comprehension and production (Frankenberg-Gacia 2014), reading (Hadley and Charles 2017), and pragmatic routines (Bardovi-Harlig *et al.* 2017).

To conclude, previous studies, both qualitative and quantitative, have shed light on the procedures and effects of learning a foreign language through corpora. While corpus use has been applied for diverse purposes in second language learning and teaching, there is still a particular pedagogical and research interest in vocabulary learning, as can be seen above. Though the majority of studies on the topic claim that the DDL approach contributes greatly to language learning, there are some limitations we should not ignore. The first and most evident problem is that most previous studies take English as the target language (Römer 2011; Vyatkina 2016), as also evidenced in a series of comprehensive surveys and syntheses (Chambers 2007; Boulton 2008, 2010b; Boulton and Cobb 2017). Consequently, the effect of DDL on other languages is inadequately tested and even remains unstudied. Secondly, according to Asención-

Delaney *et al.* (2015), there is a lack of systematicity across studies, and it is hard to make generalizations based on the findings. Lastly, some studies fail to satisfy the conventional norms of empirical research, which makes them difficult to replicate.

Given that the said challenges and issues would, to some extent, diminish the effect of DDL, more methodologically sound studies seem necessary. Therefore, the purpose of the present study is to critically and systematically examine the effect of DDL on Spanish vocabulary learning and contrast it with traditional dictionary-based activity learning. These are two research questions in the study:

1. Are DDL activities more effective than dictionary-based exercises in Spanish vocabulary learning? If so, to what extent?
2. What were the reactions of our learners of Spanish in the experimental group towards DDL activities?

## 4. METHODOLOGY

This experiment was designed to compare the efficacy of learning Spanish vocabulary through a DDL approach and a traditional dictionary-based approach after a longitudinal observation. To this end, two groups of students were involved in the study: one group dealt with paper-based DDL materials (hands-off DDL) while the other group worked with dictionary-based materials. Over the three-week experiment, the subjects were asked to participate successively in the pre-test, the immediate post-test, and the delayed post-test. A questionnaire was targeted at assessing how learners in the first group perceived DDL. All the collected data underwent a quantitative and qualitative analysis.

### *4.1. Participants*

A total of 34 conveniently available university students were recruited for this experiment, but only 32 of them completed the whole procedure. All participants were in the 20–23 age range. Most of them were female (30) while there were only two males. For the sake of comparison, 32 participants were divided into two groups: the experimental group and the control group (henceforth referred to as EG and CG respectively). Each group was made up of 16 subjects.

All participants had a similar language background and foreign language learning experience. They were all L1 Chinese undergraduates majoring in Spanish Language and Literature. They had received two or three years of formal instruction in Spanish. In general, their Spanish proficiency level was upper-intermediate.[3] All of them reported that they had not had any prior experience on corpus work.

### *4.2. Materials*

In order to analyze if vocabulary learning through DDL is more effective than learning it through dictionary consultation, a total of 38 Spanish words were selected as candidates for target items. The 38 words were extracted from the *Curricular Plan for Advanced Courses of the Specialty of Spanish* and considered advanced vocabulary, which was supposed to be unfamiliar to our participants.[4] After carrying out a pre-test in order to remove items already known by the participants, only 10 of the 38 initial words were used, namely, *clandestino, esporádico/ca, exponencial, inverosímil, latente, palpable, proliferación, rehusar, tajante,* and *vehemente* (cf. Section 4.3 for details).

Further, 30 words grouped in pairs of three were selected in order to examine if DDL is more effective than traditional learning methods, i.e. dictionary use, in allowing a distinction between confusing synonyms. The words in question were *someter, obligar, imponer, optativo, opcional, selectivo, énfasis, hincapié, relieve, proveer, suministrar, proporcionar, asignar, designar, resignar, hostil, adverso, opuesto, intenso, intensivo, tenso, colmar, apartar, involucrar, diametralmente, integralmente, sospechosamente, respetuoso/sa, respetable,* and *respetado*. Their selection was based on the suggestion, by several experienced teachers of Spanish, that these were still problematic for students of Spanish despite being familiar with them.

The learning materials consisted of two worksheets covering the same 20 target items for both groups but with different contents. Materials for the EG were based on concordance lines extracted from the annotated version of the *Corpus de Referencia del Español Actual* (CREA). The concordance lines were selected carefully according to Gilquin and Granger's (2010: 362) criteria: 'readability', 'frequency', and 'usefulness'.

[3] We chose upper-intermediate learners as participants since they were believed to have sufficient language proficiency to read high-level learning materials (such as corpus concordance and monolingual dictionaries) without the need to consult other reference books.

[4] This curricular plan is used to guide teachers and students of Spanish through the teaching and learning of Spanish in advanced courses.

Therefore, all selected concordance lines tried to avoid cut-off sentences as much as possible in order to "enhance familiarity and comprehensibility" (Moon and Oh 2017: 6). Besides, only frequent and common usages (i.e., collocation, colligation) of the target items were included so that learners could guess the word meaning and find the collocational patterns (Sripicharn 2003). To reduce participants' reading burden, manageable quantities of lines (3–5 lines in our case) were more appropriate (cf. Cobb 1997). A DDL learning material sample is given below (Figure 2).

**Infringir**

1. ¿Qué pasa en este país para que se mate, se viole, se robe y se **<infrinjan>** las leyes con el salvajismo que estamos padeciendo? ¿Qué ocurre en la juventud?

2. ¿Le pongo algunos ejemplos? El delincuente habitual es un hombre que ha decidido **<infringir>** las leyes para vivir. Comprende la necesidad de las leyes, no las discute, pero se las salta.

3. Notificamos al administrador del sitio de que estaban **<infringiendo>** las leyes electorales de California y que era necesario parar la actividad del sitio Web,

4. si es que el fiscal determina, como lo hizo con los otros dos detenidos, que **<infringieron>** la ley y por lo tanto pueden ser incluso sentenciados.

5. Si se trata de señalar a los muchachos cuando usan pelo largo, arete o drogas, de alcoholizarse, de **<infringir>** normas y causar problemas, tendríamos primero que cuestionarnos a nosotros mismos

(*fuente: Corpus de Referencia del Español Actual*)

Figure 2: Sample of DDL learning materials

Dictionaries usually offer a distinct point of comparison (Cobb 1997; Yoon and Hirvela 2004; Boulton 2010a; Anani-Sarab and Kardoust 2014, Karras 2016) since they are one of the most common resources of foreign language learning and teaching. In the worksheet of the CG, dictionary definitions or example sentences were taken from two authoritative Spanish monolingual dictionaries: 1) *Diccionario de Uso del Español* by Moliner (2007), and 2) *Diccionario de la Lengua Española* by the Royal Spanish Academy. Note that for those polysemous target items, only the dictionary meanings that matched the learning materials of DDL group were used. A sample of dictionary-based learning materials is offered below (Figure 3).

**infringir**

Infringir la ley.

Infringir las disposiciones sobre abastos.

(*fuente: Diccionario de uso del español*)

Figure 3: Sample of dictionary-based learning material

### *4.3. Instruments*

All test instruments were identical for all participants; namely, a pre-test, an immediate post-test, and a delayed post-test, except that the learning materials for the two groups in the immediate post-test were different, as pointed out before. Thus, any difference that the tests would produce can be attributed to the aid of corpus resources (Cobb 1997).

The aim of the pre-test was twofold: 1) to discard advanced words that were familiar to the subjects (cf. Anani-Sarab and Kardoust 2014; Yılmaz and Soruç 2015), and 2) to prove the homogeneity of the two groups in terms of language competence before the post-tests (Johns *et al.* 2008). For the first aim, all participants were asked to take a small and straightforward word recognition test. They had to select the words they recognized and give their corresponding definitions in L1 or L2. Note that some non-target words were added to the pre-test to disguise the target words (cf. Yılmaz and Soruç 2015). For the second aim, the scores obtained by the participants in the EEE-4 were used since they were conveniently available.[5]

The immediate post-test aimed to examine the performance of two groups after using differential learning materials. The test consisted of two parts and each part contained 10 familiar multiple-choice questions that covered all target items, with different learning materials displayed. Questions in the first part had been adapted directly from Cobb's (1997) work while the last ten questions were inspired by Boulton's (2010a) study. Each question was worth 5 points. Correct answers for each scored 5 points and incorrect answers, in turn, 0 points. The maximum score of the immediate post-test was thus 100.

Regarding the delayed post-test, its implementation was mainly targeted at comparing the effects of two differential treatments on the retention of vocabulary knowledge over a long period. The design of this test was practically identical to the

[5] EEE-4 is a Chinese national exam for undergraduate students learning Spanish, which is usually organized during the final semester of the second academic year.

former one, except that no learning materials were provided to our participants this time. Thus, the results obtained from the test were based entirely on the recall of those target items that the participants learned from the last test. Another noteworthy aspect is that the grading system for the first part of the delayed post-test was different from the previous one. The first ten questions were scored using an adapted version of the Vocabulary Knowledge Scale (Wesche and Paribakht 1996), which was a self-report scale that allowed our participants to assess how well they knew the items. Specifically, there were three scales and each scale represented different scores (cf. Table 1). In the last ten questions, the format and grading were the same as before. However, the order of questions was altered to avoid practice effects (cf. Anani-Sarab and Kardoust 2014).

| **Scale (translated from the Chinese version)** | **Score** |
|---|---|
| I know this word and I'm sure it means_____ | 5 |
| I know this word but I'm not sure what it means. It could mean_____ | 2 |
| I've seen this word before, but I don't know what it means. | 0 |

Table 1: Adapted vocabulary knowledge scale for the first ten questions of the delayed post-test

After the immediate post-test, the participants of the EG were asked to fill out a questionnaire about their perceptions of the previously performed test. The questionnaire, written in L1, consisted of two types of questions. The first one consisted of ten questions with a 5-point Likert scale. With this scale, respondents were requested to rate each question according to their agreement or satisfaction. After they had answered, scores were assigned to each response. For instance, strongly agree/very satisfied = 5; agree/satisfied = 4; neither agree/satisfied nor disagree/dissatisfied = 3; disagree/dissatisfied = 2; strongly disagree/very dissatisfied = 1. Note that to compensate for the acquiescence response bias (Lavrakas 2008), i.e., a tendency for people to agree rather than disagree in statements, some items in the questionnaire were purposely constructed, in other words, negatively worded questions, such as item 5 and 9. For example, item 9 is translated as "Do you think reading concordance lines wasted your time?" The second type was made up of five open-ended questions, in which our respondents had a chance to comment on the advantages and weaknesses of the DDL

approach. Some question samples are translated as follows: “In addition to the understanding of words’ meaning, what else did you learn from the concordance lines?” or “In comparison with traditional dictionary-based learning strategy, what are the advantages and shortcomings of concordance-based learning strategy?”

### *4.4. Procedure*

Due to limited resources and logistic problems, the experiment could not be carried out in an ordinary classroom. All test materials were posted on an online platform.[6] Admittedly, it is hard to reach the same conditions as normative tests do, but this form had several advantages: 1) subjects can freely access to the tests regardless of their location, 2) participants may not experience the same pressure as under examination conditions, and 3) the online platform can automatically grade the test and collect valuable data such as the time that each participant spends on the tests.

Before each test, a reminder was given to all participants. Firstly, the aim of the experiment was to test whether they could infer the meaning of those unfamiliar words and distinguish the synonyms based on the provided learning materials. Secondly, during the test consulting any reference books (e.g., dictionary, textbook) or online tools was not allowed; thirdly, there was no time limit for the test so they had sufficient time to read the materials and finally choose the most appropriate answers according to their level of knowledge.

The pre-test was performed once the 32 subjects had been recruited. The primary purpose of this test, as mentioned before, was to select the words that our participants could not recognize. In the end, 10 words that fulfilled the requirement were chosen. Having done this, these 10 target items, along with 10 other pairs of synonyms, were used to design materials for the later post-tests.

The immediate post-test was conducted one week after the first test. Before the participants started to answer the questions, two practice items were added so that they could familiarize themselves with the two different types of questions (cf. Stevens 1991). There were three short pauses over the test with the purpose of mitigating the fatigue effects. When the immediate post-test finished, the 16 subjects of the EG were asked to complete the questionnaire.

[6] Wenjuanxing: https://www.wjx.cn/

The last test was performed two weeks later. The participants were not informed about the form of the test so that they were not able to prepare for it or consult dictionaries beforehand. This time no practice items were provided since they were already familiar with the question types. Test breaks, however, remained unchanged.

## 5. RESULTS

### *5.1. Test results*

The first step was to ensure the homogeneity of the two groups before two post-tests. As mentioned before, the scores that our participants obtained from the EEE-4 were conveniently utilized. The descriptive statistics of the scores roughly indicated similar language competence between the EG and the CG ($M_{EG}$ = 79.56, $SD_{EG}$ = 6.18; $M_{CG}$ = 77.31, $SD_{CG}$ = 6.77). A later *t*-test confirmed that the two groups did not differ significantly ($p$ = .334), thereby suggesting that any diverging outcome would be due to different experimental treatments, i.e., DDL approach and traditional method.

Based on the data collected in the immediate post-test, it can be clearly noticed that the EG ($M$ = 78.75, $SD$ = 12.58) achieved greater performance than the CG ($M$ = 50.00, $SD$ = 13.66). However, this did not show whether the difference between them was statistically significant or not. An independent samples *t*-test based on groups' mean score seemed appropriate. But before testing this hypothesis, it was necessary to check whether the collected data satisfies the assumptions of the *t*-test, namely, ñthe normality assumptionò (Brezina 2018: 13). The type of normality tests to be chosen depends on the sample size. If the sample size is big, Kolmogorov-Smirnov test is used and, if it is not, Shapiro-Wilk test is used instead.[7] Given that our sample size is small ($N$ = 32), the Shapiro-Wilk normality test was run. The result demonstrated that the data of the two groups did conform to a normal distribution ($p_{EG}$ = .078, $p_{CG}$ = .577). The next step was to carry out the *t*-test, whose result (cf. Table 2) indicated that there was a statistically significant difference between the DDL group and the dictionary-based group in the immediate test [$t(30) = 6.191$, $p < .001$, 95% CI (19.27, 38.23)]. However, what the *t*-test was unable to tell is "how large this difference is and whether it is

[7] There is no accurate standard reference regarding if the sample size is big enough for a Kolmogorov-Smirnov test, but the conventional cut-off size is 50 (https://statistics.laerd.com/spss-tutorials/testing-for-normality-using-spss-statistics.php). Recent research, however, shows that *t*-test can be robust to the violation of the normality assumption, which means that even if the data follow an abnormal distribution that does not interfere with valid results of the *t*-test (cf. Brezina 2018).

practically important" (Brezina 2018: 14). Therefore, it is necessary to report effect sizes here. In our case, the well-known Cohen's *d* was used to measure the effect size of the difference. The result ($d = 2.19$) suggested that it was much larger than Plonsky and Oswald's (2014) L2 field-specific criterion for a large effect size ($d = 1.00$).

| **Groups** | ***N*** | ***M*** | ***SD*** | ***t*-value** | **df** | ***p*** |
|---|---|---|---|---|---|---|
| Experimental group | 16 | 78.75 | 12.58 | 6.191 | 30 | < .001 |
| Control group | 16 | 50.00 | 13.66 | | | |

Table 2: Result of *t*-test for the two groups in the immediate post-test

The same statistical tests were then performed in the delayed post-test. First, the Shapiro-Wilk normality test showed that the data collected from the two groups were also normally distributed ($p_{EG} = .364$, $p_{CG} = .069$). Subsequently, the results of the *t*-test are presented in Table 3. As can be noticed, there was a statistically significant difference between the two groups [$t(30) = 2.600$, $p = .014$, 95% CI (3.54, 29.46)]. Moreover, the effect size of this difference was $d = 0.92$, which was close to the L2 field-specific benchmark for a large size (cf. Plonsky and Oswald 2014).

| **Groups** | ***N*** | ***M*** | ***SD*** | ***t*-value** | **df** | ***p*** |
|---|---|---|---|---|---|---|
| Experimental group | 16 | 49.19 | 18.29 | 2.600 | 30 | .014 |
| Control group | 16 | 32.69 | 17.61 | | | |

Table 3: Result of *t*-test for the two groups in the delayed post-test

Based on the results of the between-group comparisons in two post-tests, the positive effect of DDL approach on Spanish vocabulary learning is quite clear and statistically meaningful. It seems that learners in the EG performed better in terms of lexical awareness and word recall with the aid of corpus resources, i.e., concordance-based materials. The CG, in contrast, did not achieve satisfying results using only dictionary-based materials. Nevertheless, it is worth noticing that there was a marked decrease in the mean score of both groups between the two post-tests, as illustrated in Table 4.

| | Experimental group | Control group |
|---|---|---|
| **Immediate post-test** | 78.75 | 50.00 |
| **Delayed post-test** | 49.19 | 32.69 |
| **Difference** | -29.56 | -17.31 |
| **Change (% of the immediate post-test)** | -37.54% | -34.62% |

Table 4: Result of within-group comparison between the immediate post-test and delayed post-test

As can be seen in Table 4 above, both groups experienced a decrease of nearly 35% in the mean score, which means that all participants had difficulties in recalling words learned two weeks before. Further examination revealed that the first part of the test materials (i.e., advanced vocabulary) contributed to such a significant decrease: the average score in this part declined by 57.48% (-55.90% for the EG and -59.06% for the CG). In the synonym discrimination part, however, some students showed better performance in the delayed post-test than the immediate one. Possible explanations for this will be discussed in due course.

### *5.2. Questionnaire results*

The ten Likert-type questions were analyzed first. According to their attitudes and beliefs, 15 out of 16 participants were generally satisfied with the concordance-based activity, 13 students liked the sentences displayed in KWIC format and 15 agreed that the concordance lines were readily comprehensible. All participants agreed that concordance lines provided a rich context for the target vocabulary and 15 considered that concordance lines helped them differentiate those synonyms. The encouraging feedback from the participants was enhanced when they were asked whether they would support the application of corpus resources to Spanish learning: 14 participants gave an affirmative answer.

Using a 5-point Likert scale, participants' perceptions were quantifiable. The mean score and standard deviation for each question are displayed below (cf. Table 5).

| Item | Mean | *SD* |
|---|---|---|
| 1 | 4.37 | 0.619 |
| 2 | 4.13 | 0.806 |
| 3 | 4.25 | 0.775 |
| 4 | 4.31 | 0.602 |
| 5 | 2.94 | 1.340 |
| 6 | 4.56 | 0.512 |
| 7 | 4.44 | 0.629 |
| 8 | 4.62 | 0.719 |
| 9 | 2.50 | 1.155 |
| 10 | 4.44 | 0.727 |

Table 5: Descriptive statistics for the first ten Likert-type questions

Overall, our participants rated the concordance-based activities highly, except for item 5 and 9. As mentioned in Section 4.3, these two items were asked from a negative perspective of DDL. A larger standard deviation of the two items suggested an evident variation in our participants' perception. In other words, some students indeed thought reading concordance lines wasted their time while others did not agree.

Regarding the highly-rated scores, if these questions *per se* were poorly designed, the result would not be as reliable as it looked. Therefore, Cronbach's alpha was utilized to measure the internal consistency estimate of the reliability of the rated scores. The first step was to eliminate item 5 and 9 since they were out of line with the other eight questions. After this, the reliability analysis was performed. The result is provided in Table 6: it indicates that the coefficient alpha for the eight *pro-DDL* questions is statistically acceptable ($\alpha = .74$), and it confirms that the majority of our respondents had a positive attitude towards DDL.[8]

[8] According to the rules of thumb, $0.7 \leq \alpha < 0.8$ means the internal consistency is acceptable.

| Cronbach's alpha | Number of items |
|---|---|
| 0.74 | 8 |

Table 6: Cronbach's alpha reliability coefficient for 8 Likert-scale questions

With regard to the five open-ended questions, the participants had an ambivalent reaction to DDL activities. Most respondents considered the concordance lines clear, and they provided rich contexts for target vocabulary and synonym distinction. However, many expressed their dissatisfaction with the quantity and length of those concordance lines. Some unfamiliar words also affected the understanding of the whole sentence. When they were asked whether they had obtained more valuable information based on the concordance lines, 13 respondents answered yes. Among them eight mentioned that the collocations helped them recognize words and answer the questions, two said that they knew the common usage of the words thanks to concordance lines, and one pointed out that concordance lines of each synonym pair helped discriminate their meanings.

Given that all our participants had previous experience in learning Spanish through dictionary-based activities, the next question was about the advantages and shortcomings of concordance-based activities (i.e., DDL) when compared to traditional learning strategies. Half of the participants believed that they understood the meaning of the words more precisely with the help of contexts. Three of them stated they had a more profound impression of those words because of their repeated occurrences in the context. Four held the opinion that concordancing was useful in terms of synonym discrimination. Two learners gained benefit from the collocations. As regards the disadvantages, many complained about the time they spent on the test. This feeling was also evidenced by the time recorder of the online platform: the average duration of the immediate post-test for the EG was more than 30 minutes, contrasting with around 20 minutes for the CG. Four participants mentioned that the concordance lines were "too lengthy and too many" to have the patience to read them all. Another four claimed that, in this sense, the dictionary was more convenient and straightforward to consult the meaning of a word. An interesting comment by one of the participants was that "with this method [concordance-based approach], it is hard to understand the meaning of the

word, while a bilingual dictionary can provide more precise definitions and more common collocations," while another comment pointed out that "the vocabulary learned through concordance will be easily forgotten: concordance lines will only help memorize the meaning of the word and usages during a short run."

To summarize, most comments from our participants were considered positive. They generally agreed that the context was rich enough to learn words, collocations were easy to detect, and concordance lines helped the distinction of synonym pairs. At the same time, some problems of DDL should not be neglected, for instance, the time-consuming process and the interference from unfamiliar words. These drawbacks are also reported in several previous studies (cf. Cheng *et al.* 2003; Yoon and Hirvela 2004; Chambers 2005; Chambers 2007; Boulton 2010a; Geluso and Yamaguchi 2014).

## 6. Discussion and conclusions

Based on the quantitative and qualitative analyses of the two post-tests and a questionnaire assessment, DDL activities have been found to be more effective and efficient than the traditional dictionary-based activities in terms of vocabulary learning. And the difference between the two treatments was statistically significant with a large effect size (research question 1). Moreover, most of our participants took a positive attitude towards DDL activities and its future application (research question 2).

The study has provided empirical evidence for the effectiveness of DDL activities on a language other than English, which is meaningful for the popularization and acceptance of DDL. Also, the study has followed the procedures of empirical research by drawing on previous pioneering studies (e.g., Cobb 1997; Boulton 2010a; Anani-Sarab and Kardoust 2014). Hence, the research design of the present experiment is methodologically and statistically sound. Besides, the details and procedure of the experiment have been reported as much detail as possible so that it would be feasible to replicate it in the future.

Although the findings of the present study are in line with the reassuring conclusions from prior research, such as Cobb (1997, 1999), Chambers (2005), Boulton (2007, 2010a), Soruç and Tekin (2017), the experimental results of our particular case should be interpreted cautiously. The first point to note is that two learning materials in the immediate post-test (cf. Figure 2 and Figure 3) presented an unbalanced amount of

information. The DDL learning materials had more textual input than the dictionary-based ones, a feature that also characterized Cobb's (1997) experiment. In our case, since less informative definitions and example sentences are one of the major flaws of two Spanish dictionaries (cf. Calderón-Campos 1994 for discussion), the learning materials for the control group inevitably offered less textual input. The contents of the concordance-based materials, on the other hand, may be problematic as well. For the good of our participants, only a manageable number of concordance lines (3–5 lines in our case) were selected. However, the cherry-picking lines may expose a bias in favor of researchers or teachers instead of catering for learners. For example, sometimes the concordance lines may have contained vocabulary that is beyond the proficiency level of learners, but the researcher may not have realized it. With a few lines, moreover, it is easy to go to extremes: at one end, concordance provides too many similar examples of one specific usage, and learners will get bored easily and, at the other end, concordance contains too little or no data that students would like to learn, and they probably will get frustrated (cf. Flowerdew 1996).

Secondly, from the immediate post-test to the delayed post-test, both groups underwent a substantial drop in performance (a loss of nearly 35% in the mean score), a result that merits further discussion. The primary factor could be that the forgetting curve was at work. Since the interval between the two post-tests was two weeks, and during that period our subjects did not receive any similar treatments, forgetting the knowledge learned before seems a logical outcome. Another explanation for this is that the students did not receive any feedback or correction after the immediate post-test. Thus, there was no way for them to know what the correct answers were, let alone to learn from the potential errors. In this situation, the participants would answer the questions of the last post-test based merely on what they learned during the immediate post-test. Theoretically, it seems unlikely that the performance in the delayed post-test reaches the same level as in the former post-test. Indeed, except for two subjects from the EG and another three from the CG who succeeded in keeping the same scores as the previous test, the remainder obtained lower scores in the last test.

Another potential issue relates to the results emerging from the questionnaire assessment. In both Likert-type and open-ended questions, the majority of our participants gave a positive evaluation of this brand-new learning method, a result in line with many other pertinent studies (e.g., Yoon and Hirvela 2004; Boulton 2010a;

Yılmaz and Soruç 2015; Moon and Oh 2017). Undoubtedly, participants' positive reaction is highly significant for the public acceptability of DDL and its future application to foreign language learning and teaching, but "their [participants'] subjective appreciations of their own learning may not be reflected in actual learning" (Boulton 2010b: 3), because the overwhelmingly positive reactions may be ascribable to 'the novelty factor' and 'the Hawthorne effect' (Boulton 2017b: 185). In addition, the acquiescence bias (Lavrakas 2008) in those responses is hard to avoid even with negatively worded items involved (cf. Section 4.3), and this is especially true when questions belong to the agree/disagree type. In other words, participants tend to put in a good word for the formulated statements because they would like to show their politeness and respect. Another issue to address here is that the coefficient alpha for the first part of the questionnaire was statistically acceptable but did not reach the benchmark of good or excellent. The reason could be due to the small number of Likert-type questions. Typically, a larger number of items lead to a larger α.

Despite the fact that the present study aims to examine the effectiveness and efficiency of DDL, there is one caveat: its objective does not lie in demonstrating that traditional methods like dictionary-based activities are not effective in foreign language learning and teaching. In reality, as one of the most traditional and classical study tools, the dictionary is and will play a fundamental role in language learning and teaching (Anani-Sarab and Kardoust 2014). Interestingly, nowadays corpora exert considerable influence on the publishing work and has hastened the birth of corpus-based dictionaries (McEnery and Wilson 1997; Römer 2011). In this sense, the new generation of dictionaries shows the corpus to good advantage. Returning to our study, it is noteworthy that the participants of both groups did not recognize the target items at first. But with the aid of both learning materials, all of them have acquired vocabulary knowledge. However, as Chambers (2010) indicates, in working with concordance, the learner can check and confirm whether one particular use is correct or not, thereby reinforcing the learning process. The longitudinal experiment conducted by Cobb (1999), as mentioned before, also suggests that both concordance and dictionary information bring benefits in the short run, but only the concordance group retains knowledge for an extended period. From the students' point of view, DDL also outweighs dictionary learning in terms of the benefits. As Yoon and Hirvela (2004: 277) reported, "they [participants] agreed that a dictionary is useful for acquiring the

meaning of words, but a corpus is more useful for learning how and where to put words in context."

We should also acknowledge there are several obstacles on the way to implementing DDL, which is the reason DDL is still placed outside the mainstream of foreign language learning and teaching, or 'marginal practice' in Boulton's term (2017a: 483). There are several possible explanations for this situation. First of all, infrastructure or logistics is one of the biggest problems for DDL (Gilquin and Granger 2010). If it is the case of hands-on DDL, its implementation means that schools or universities have to be well equipped with computers and servers, which will increase the equipment budget (Pérez-Paredes 2005). Moreover, if corpus resources are not freely available, buying licenses also costs a large amount of money. Secondly, corpus linguistics is a relatively new field whereas the techniques of corpus are developing at a dizzy speed. Many foreign language teachers have never received any relevant training and probably are reluctant to adopt a new method like DDL (Pérez-Paredes 2005). Thus, how to make teachers "corpus literate and comfortable with mechanics of corpus analysis" (Szudarski 2018: 106) is a problem. Another issue that possibly emerges is the dilemma of the teacher's role. DDL places more emphasis on the central role of the learner during the learning process, well known as 'learner-centred' (Mukherjee 2006: 12). On the one hand, learners' autonomy reaches the maximum with DDL; teachers play a less central role and have less control over the class, on the other hand (Gilquin and Granger 2010). How to balance teacher-led against learner-led awaits more discussion. Thirdly, from learner's perspective, they need to overcome technophobia if they are to search a corpus personally. Besides, they also should ideally know some basic corpus query skills. Even with the soft version of DDL, some learners still think reading paper-based concordance is time-consuming and difficult. In this respect, Gilquin and Granger (2010) argue that DDL may only be suitable for certain learners, depending on their learning style. All these issues above are challenging the development and application of DDL. As Römer (2011: 206) pointed out, there is still "much work to be done in bridging the gap between research and practice."

The scope of this study is limited regarding the sample size and the representativeness. Only 32 subjects were involved in the experiment. All of them were from universities and with an upper-intermediate level of Spanish. Thus, the results may not apply to other Spanish learners with different proficiency levels. Future work needs

to increase the size and variety of sample. It is recommended to use power analysis (Faul *et al.* 2007) to determine the required sample size based on a fixed α, power, and effect size. More studies are needed to examine the effectiveness of DDL on other levels of language proficiency (e.g., lower level, intermediate level), especially in the SFL context. It would also be profitable to include more linguistic features in future research instead of only focusing on one or two since language learners are supposed to master a language as a whole (Boulton 2017b). Lastly, future empirical studies could extend the experiment to see whether the DDL activities are more beneficial following the continuum from teacher-led to learner-led. By making these endeavors, we have reason to believe that DDL will have a vast potential for its future application and popularization.

## REFERENCES


Allan, Rachel. 2006. *Data-driven Learning and Vocabulary: Investigating the Use of Concordances with Advanced Learners of English*. Dublin: Trinity College Dublin.

Anani-Sarab, Mohammad R. and Amir Kardoust. 2014. Concordance-based data-driven learning activities and learning English phrasal verbs in EFL classrooms. *Issues in Language Teaching* 3/1: 89–112.

Asención-Delaney, Yuly, Joseph G. Collentine, Karina Collentine, Jersus Colmenares and Luke Plonsky. El potencial de la enseñanza del vocabulario basada en corpus: Optimismo con precaución. *Journal of Spanish Language Teaching* 2/2: 140–151.

Aşık, Asuman, Arzu Sarlanoglu Vural and Kadriye Dilek Akpınar. 2016. Lexical awareness and development through data driven learning: Attitudes and beliefs of EFL learners. *Journal of Education and Training Studies* 4/3: 87–96.

Barcroft, Joe. 2005. La enseñanza del vocabulario en español como segunda lengua. *Hispania* 88/3: 568–582.

Bardovi-Harlig, Kathleen, Sabrina Mossman and Yunwen Su. 2017. The effect of corpus-based instruction on pragmatic routines. *Language Learning and Technology* 21/3: 76–103.

Bernardini, Silvia. 2000. Systematising serendipity: Proposals for concordancing large corpora with language learners. In Lou Burnard and Tony McEnery eds. *Rethinking Language Pedagogy from a Corpus Perspective*. Hamburg: Peter Lang, 225–234.

Bernardini, Silvia. 2004. Corpora in the classroom: An overview and some reflections on future developments. In John Sinclair ed. *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins, 5–36.

Boulton, Alex. 2007. But where's the proof? The need for empirical evidence for data-driven learning. In Michael Edwardes ed. *Proceedings of the BAAL Annual Conference 2007*. London: Scitsiugnil Press, 13–16.

Boulton, Alex. 2008. Esprit de corpus: Promouvoir l'exploitation de corpus en apprentissage des langues. *Texte et Corpus* 3: 37–46.

Boulton, Alex. 2009. Testing the limits of data-driven learning: Language proficiency and training. *ReCALL* 21/1: 37–54.

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

Boulton, Alex. 2010b. Learning outcomes from corpus consultation. In María Moreno-Jaén, Fernando Serrano Valverde and María Calzada Pérez eds. *Exploring New Paths in Language Pedagogy: Lexis and Corpus-Based Language Teaching*. London: Equinox, 129–144.

Boulton, Alex. 2010c. Data-driven learning: On paper, in practice. In Tony Harris and María Moreno-Jaén eds. *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, 17–52.

Boulton, Alex. 2017a. Research timeline: Corpora in language teaching and learning. *Language Teaching* 50/4: 483–506.

Boulton, Alex. 2017b. Data-driven Learning and Language Pedagogy. In Steven L. Thorne and Stephen May eds. *Language, Education and Technology*. New York: Springer, 181–192.

Boulton, Alex and Tom Cobb. 2017. Corpus use in language learning: A meta-analysis. *Language Learning* 67/2: 1–46.

Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Calderón-Campos, Miguel. 1994. Sobre la elaboración de diccionarios monolingües de producción: Las definiciones, los ejemplos y las colocaciones léxicas. In Peter Jan Slagter ed. *Aproximaciones a Cuestiones de Adquisición y Aprendizaje del Español como Lengua Extranjera o Lengua Segunda*. Amsterdam: Rodopi, 105–119.

Chambers, Angela and Íde O'Sullivan. 2004. Corpus consultation and advanced learners' writing skills in French. *ReCALL* 16/1: 158–172.

Chambers, Angela. 2005. Integrating corpus consultation in language studies. *Language Learning and Technology* 9/2: 111–125.

Chambers, Angela. 2007. Popularising corpus consultation by language learners and teachers. In Encarnación Hidalgo, Luis Quereda and Juan Santana eds. *Corpora in the Foreign Language Classroom*. Amsterdam: Rodopi, 3–16.

Chambers, Angela. 2010. What is data-driven learning? In Anne O'Keeffe and Michael McCarthy eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 345–358.

Chan, Tun-pei and Hsien-Chin Liou. 2005. Effects of web-based concordancing instruction on EFL students' learning of verb-noun collocations. *Computer Assisted Language Learning* 18/3: 231–251.

Chang, Ji-Yeon. 2014. The use of general and specialized corpora as reference sources for academic English writing: A case study. *ReCALL* 26/2: 243–259.

Chang, Jung. 2001. Chinese speakers. In Michael Swan and Bernard Smith eds. *Learner English: A Teacher's Guide to Interference and other Problems*. Cambridge: Cambridge University Press, 310–324.

Charles, Maggie. 2012. Proper vocabulary and juicy collocations: EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes* 31/2: 93–102.

Cheng, Winnie, Martin Warren and Xun-feng, Xu. 2003. The language learner as language researcher: Putting corpus linguistics on the timetable. *System* 31/2: 173–186.

Cobb, Tom. 1997. Is there any measurable learning from hands-on concordancing? *System* 25/3: 301–315.
Cobb, Tom. 1999. Breadth and depth of lexical acquisition with hands-on concordancing. *Computer Assisted Language Learning* 12/4: 345–360.
Cotos, Elena, Stephanie Link and Sarah Huffman. 2017. Effects of DDL technology on genre learning. *Language Learning and Technology* 21/3: 104–130.
Crosthwaite, Peter. 2017. Retesting the limits of data-driven learning: Feedback and error correction. *Computer Assisted Language Learning* 30/6: 447–473.
Daskalovska, Nina. 2015. Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning* 28/2: 130–144.
Faul, Franz, Edgar Erdfelder, Albert-Georg Lang and Axel Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39/2: 175–191.
Flowerdew, John. 1996. Concordancing in language learning. In Martha C. Pennington ed. *The Power of CALL*. Houston: Athelstan, 97–113.
Frankenberg-Garcia, Ana. 2014. The use of corpus examples for language comprehension and production. *ReCALL* 26/2: 128–146.
Gabrielatos, Costas. 2005. Corpora and language teaching: Just a fling or wedding bells? *The Electronic Journal for Teaching English* 8/4: 1–32.
Gaskell, Delian and Thomas Cobb. 2004. Can learners use concordance feedback for writing errors? *System* 32/3: 301–319.
Gavioli, Laura. 2001. The learner as researcher: Introducing corpus concordancing in the classroom. In Guy Aston ed. *Learning with Corpora*. Bologna: Athelstan, 108–137.
Geluso, Joe and Atsumi Yamaguchi. 2014. Discovering formulaic language through data-driven learning: Student attitudes and efficacy. *ReCALL* 26/2: 225–242.
Gilmore, Alex. 2009. Using online corpora to develop students' writing skills. *ELT Journal* 63/4: 363–372.
Gilquin, Gaëtanelle and Sylviane Granger. 2010. How can data-driven learning be used in language teaching? In Anne O'Keeffe and Michael McCarthy eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 359–370.
Guan, Xiaowei. 2013. A study on the application of data-driven learning in vocabulary teaching and learning in China's EFL class. *Journal of Language Teaching and Research* 4/1: 105–112.
Hadley, Gregory and Maggie Charles. 2017. Enhancing extensive reading with data-driven learning. *Language Learning and Technology* 21/3: 131–152.
Huang, Zeping. 2014. The effects of paper-based DDL on the acquisition of lexico-grammatical patterns in L2 writing. *ReCALL* 26/2: 163–183.
Jiao, Binkai. 2012. An empirical study on corpus-driven English vocabulary learning in China. *English Language Teaching* 5/4: 131–137.
Jiménez-Calderón, Francisco and Ana Sánchez-Rufat. 2017. Posibilidades de aplicación de un enfoque léxico a la enseñanza comunicativa del español. In Guadalupe Nieto Caballero ed. *Nuevas Aportaciones al Estudio de la Enseñanza y Aprendizaje de Lenguas*. Cáceres: Universidad de Extremadura, 11–23.
Johns, Tim. 1991. Should you be persuaded: Two examples of data-driven learning materials. *ELR Journal* 4: 1–16.
Johns, Tim, Lee Hsingchin and Wang Lixun. 2008. Integrating corpus-based CALL programs in teaching English through children's literature. *Computer Assisted Language Learning* 21/5: 483–506.

Karras, Jacob N. 2016. The effects of data-driven learning upon vocabulary acquisition for secondary international school students in Vietnam. *ReCALL* 28/2: 166–186.

Kennedy, Claire and Tiziana Miceli. 2010. Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource. *Language Learning and Technology* 14/1: 28–44.

Koosha, Mansour and Ali A. Jafarpour. 2006. Data-driven learning and teaching collocation of prepositions: The case of Iranian EFL adult learners. *Asian EFL Journal* 8/4: 192–209.

Lavrakas, Paul J. 2008. *Encyclopedia of Survey Research Methods*. California: SAGE Publications.

Lee, Hansol, Mark Warschauer and Jang H. Lee. 2019. The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics* 40/5: 721–753.

Lewis, Michael. 1993. *The Lexical Approach: The State of ELT and a Way Forward*. Boston: Heinle.

Li, Shuangling. 2017. Using corpora to develop learners' collocational competence. *Language Learning and Technology* 21/3: 153–171.

McEnery, Tony and Andrew Wilson. 1997. Teaching and language corpora. *ReCALL* 9/1: 5–14.

Mizumoto, Atsushi and Kiyomi Chujo. 2015. A meta-analysis of data-driven learning approach in the Japanese EFL classroom. *English Corpus Studies* 22: 1–18.

Moliner, María. 2007. *Diccionario del Uso del Español*. Gredos: Madrid.

Moon, Soyeon and Sun-Young Oh. 2017. Unlearning overgenerated *be* through data-driven learning in the secondary EFL classroom. *ReCALL* 30/1: 48–67.

Mukherjee, Joybrato. 2006. Corpus linguistics and language pedagogy: The state of the art – and beyond. In Sabine Braun, Kurt Kohn and Joybrato Mukherjee eds. *Corpus Technology and Language Pedagogy*. Bern: Peter Lang, 5–24.

Nation, Paul and Paul Meara. 2010. Vocabulary. In Norbert Schmitt ed. *An Introduction to Applied Linguistics*. London: Hodder Education, 34–51.

Nation, Paul. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

O'Keeffe, Anne, Michael McCarthy and Ronald Carter. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

O'Sullivan, Íde and Angela Chambers. 2006. Learners' writing skills in French: Corpus consultation and learner evaluation. *Journal of Second Language Writing* 15/1: 49–68.

Pérez-Paredes, Pascual. 2005. Data-driven learning y el aprendizaje de idiomas. *Greta: Revista para Profesores de Inglés* 13/1–2: 5–10.

Pérez-Paredes, Pascual. 2010. Appropriation and integration issues in corpus methods and mainstream language education. In Tony Harris and María Moreno-Jaén eds. *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, 53–73.

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

Römer, Ute. 2011. Corpus research applications in second language teaching. *Annual Review of Applied Linguistics* 31: 205–225.

Royal Spanish Academy. Online. *Diccionario de la Lengua Española* (22. ed.). At https://dle.rae.es/?w=diccionario. Accessed on 25/09/2019

Royal Spanish Academy. Online. Banco de Datos CREA: *Corpus de Referencia del Español Actual*. At http://www.rae.es. Accessed on 25/09/2019

Schmitt, Norbert. 2000. *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.
Smart, Jonathan. 2014. The role of guided induction in paper-based data-driven learning. *ReCALL* 26/2: 184–201.
Soruç, Adem and Bilal Tekin. 2017. Vocabulary learning through data-driven learning in an English as a second language setting. *Educational Sciences: Theory and Practice* 17/6: 1811–1832.
Sripicharn, Passapong. 2003. Evaluating classroom concordancing: The use of concordance-based materials by a group of Thai students. *Thammasat Review* 1: 203–236.
Sripicharn, Passapong. 2010. How can we prepare learners for using language corpora? In Anne O'Keeffe and Michael McCarthy eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 371–384.
Stevens, Vance. 1991. Concordance-based vocabulary exercises: A viable alternative to gap-filling. *English Language Research Journal* 4: 47–61.
Szudarski, Paweł. 2018. *Corpus Linguistics for Vocabulary: A Guide for Research*. London: Routledge.
Tekin, Bilal and Adem Soruç. 2016. Using corpus-assisted learning activities to assist vocabulary development in English. *The Turkish Online Journal of Educational Technology* 1270–1283.
Thompson, Paul. 2006. Assessing the contribution of corpora to EAP practice. In Zoe Kantaridou, Iris Papadopoulou and Ifigenia Mahili eds. *Motivation in Learning Language for Specific and Academic Purposes*. Macedonia: University of Macedonia.
Thurstun, Jennifer and Christopher N. Candlin. 1998. Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes* 17/3: 267–280.
Vyatkina, Nina. 2016. Data-driven learning for beginners: The case of German verb-preposition collocations. *ReCALL* 28/2: 207–226.
Wesche, Marjorie and T. Sima Paribakht. 1996. Assessing Second language vocabulary knowledge: Depth versus breadth. *The Canadian Modern Language Review* 53/1: 13–40.
Yılmaz Enes and Adem Soruç. 2015. The use of concordance for teaching vocabulary: A data-driven learning approach. *Procedia-Social and Behavioral Sciences* 191: 2626–2630.
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*
Gang Yao
University of Murcia
Department of English Philology
30001 Calle Santo Cristo 1
Spain
e-mail: gang.yao@um.es

received: October 2018
accepted: September 2019

RiCL Research in Corpus Linguistics

# ‘A matter both of curioſity and uſefulneſs’: Compiling the *Corpus of English Texts on Language*

Leida Maria Monaco - Luis Puente-Castelo
University of Oviedo & University of A Coruña / Spain

**Abstract** – This paper describes the compilation of CETeL, the subcorpus on ‘Language and Linguistics’ in the *Coruña Corpus of English Scientific Writing*, and discusses the various challenges encountered during the process of selection and digitisation of material. CETeL includes forty-four samples of texts on Language, Languages, and Linguistics from the period 1700–1900, and on completion will contain around 400,000 words. The paper will examine the historical context of academic writing in that period and the way in which this context affects the process of compilation. Likewise, the criteria followed in the compilation of the *Coruña Corpus* will be discussed in order to show the extent to which these criteria have affected the compilation of CETeL, and how they contribute towards making the corpus representative of the disciplinary practices of the period. Finally, the corpus will also be described according to a series of parameters used to assure representativeness and balance, namely the date of publication of samples, their genre, and the sex and linguistic background of their authors.

**Keywords** – *Coruña Corpus*; corpus compilation; Late Modern English; scientific writing

## 1. INTRODUCTION[1]

The *Corpus of English Texts on Language* (henceforth, CETeL) is one of the many twin subcorpora of the *Coruña Corpus of English Scientific Writing*, currently under compilation at the Universidade da Coruña by the Research Group for Multidimensional Corpus-Based Studies in English (MuStE, http://www.udc.es/grupos/muste). This paper covers the process of compilation and selection of samples in CETeL, which has now been completed,[2] and discusses the challenges faced here, focusing in particular on the

[1] The research reported here has been funded by the Spanish Ministry of Economy and Competitiveness (MINECO), grant number FFI2016-75599-P. This grant is hereby gratefully acknowledged.
[2] The initial process of computerisation of CETeL is complete, and a process of revision is about to start.

difficult task of collecting a set of samples sufficiently representative of the type of language used in writing about language between 1700 and 1900, and on how these challenges were approached.

Section 2 presents the general design of the *Coruña Corpus*, to which CETeL belongs, while Section 3 explains its general compilation criteria. The *status quo* of Language and Linguistics studies in the eighteenth and nineteenth centuries is dealt with in Section 4, and Section 5 provides an analysis of the difficulties in reconciling general criteria and disciplinary particularities during the compilation of CETeL. Finally, a thorough description of CETeL is offered in Section 6, looking at a series of parameters including the distribution of samples over time, their topics, their genres, and the sex and linguistic background of their authors, followed by brief concluding remarks in Section 7.

## 2. The *Coruña Corpus*

Designed to be a "purpose-built electronic corpus conceived of as a resource for the study of scientific writing in English" (cf. Moskowich 2012: 35), the *Coruña Corpus* contains samples of texts of a scientific nature from the eighteenth and nineteenth centuries, allowing research at all linguistic levels except phonology. The corpus will consist of ten subcorpora (see Figure 1), all with the same design and principles of compilation, and one for each field of knowledge or scientific discipline.

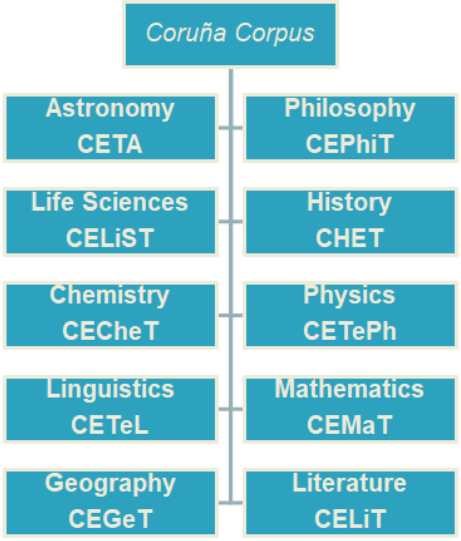


Figure 1: Current plan for subcorpora in the *Coruña Corpus*

Three of these ten subcorpora – on Astronomy (CETA), Philosophy (CEPhiT), and History (CHET) – have already been published. The former two were originally released along with collections of pilot studies (cf. Moskowich and Crespo 2012 and Moskowich *et al.* 2016, respectively).[3] An edited volume with pilot studies involving CHET is also in the final stages of preparation (cf. Moskowich *et al*. 2019).

In addition, three other subcorpora – on Life Sciences (CELiST), Chemistry (CECheT), and Linguistics (CETeL) – are under compilation, each currently at a different level of completion, whereas the subcorpora on Mathematics, Physics, Literature, and Geography are still at very early stages of development.

### *2.1. Size*

As noted above, all the subcorpora of the *Coruña Corpus* have the same design and structure. Each subcorpus contains a series of samples of approximately 10,000 words in length, at a rate of two samples per decade, leading to a total of 20,000 words per decade and discipline, and hence 200,000 words per century and per discipline, and 400,000 words per subcorpus. This has been done with a view to making the *Coruña Corpus* approximately 4,000,000 words long when completed, a size which, arguably, should allow enough variety of texts and genres to dilute the influence of any idiosyncrasies in particular texts and to make the corpus representative of the scientific writing of the period.

The size of the samples has also been a matter of conscious selection, and the number of approximately 10,000 words is far from arbitrary. Despite the fact that Biber (1993) has argued in favour of samples as small as 1,000 words long, the compilers of the *Coruña Corpus* took into consideration the fact that the scientific register was less standardised between 1700 and 1900 than it is today, and hence the possibility that 1,000-word samples might not provide a good representation of the register during this period.

A further problem with such small samples is that they would inevitably slow down the process of compilation due to the difficulty in attaining a corpus of an adequate overall size, given the limited number of valid texts available for inclusion. In

[3] All three are now freely available in open access form at the Universidade da Coruña Open Repository at https://ruc.udc.es/dspace/handle/2183/21846.

terms of content, samples are usually selected in such a way that they cover all parts and sections of texts (such as introductions, methods, results, discussions, or/and conclusions, but as a general rule excluding prefaces), in order to avoid accusations of arbitrariness such as those mentioned by Claridge *et al*. (1999: introduction), who see text samples as "arbitrarily cut-out smaller text chunks" and suggest that full texts should be selected instead.

It is worth noticing that 'register' is understood here in Biber and Conrad's (2009) sense, namely as a variety of language characterised by having particular communicative purposes for particular situations (i.e. scientific texts). We also understand register as a scalable concept, which can be "defined at varying levels of specificity" (cf. Gray 2011: 3), as registers can be influenced by several situational factors at once. Thus, in our study, texts on language are considered a subregister of scientific texts on account of the particular constraints of a discipline, and texts by women are considered a subregister on account of the particular constraints faced by women authors. We refer to these types of situational factors as 'parameters' (sex, linguistic background, discipline, etc.), as they are both used to assure the representativeness of the selection of texts and as possible parameters for study, and the possible values these parameters have as 'categories' (female, Irish, linguistics). By contrast, 'genre' is understood here as a recurrent formal structure adopted by a variety of language as a result of conventions on how information is organised formally in order to achieve a given purpose (i.e. a research article). Thus, it is considered as one of the above-explained parameters, accounting for the particular constraints posed by formats and formal issues.

### *2.2. Timespan*

The *Coruña Corpus* contains samples of scientific writing from the eighteenth and nineteenth centuries, a period of profound change both in science and in the way science was written (cf. Beal 2012). This period is delimited by two important events which might be considered as chronological bookends.

The early eighteenth century marks the culmination of the process of change in science which had begun in the seventeenth century with the works by Francis Bacon and Boyle and which saw scholasticism being replaced by a new scientific paradigm

(cf. Taavitsainen and Pahta 1998: 162). This coincides with the dissemination of Newton's ideas on gravity, which revolutionised the understanding and practice of physics and would go on to influence a great deal of scientific research over the following two centuries. The first years of the twentieth century, in turn, coincide with several major scientific breakthroughs, perhaps the most important being Einstein's 1905 paper on the Special Theory of Relativity, which is still considered a foundation for research in many disciplines.

The period in between these turning points is one of constant innovations and, at the linguistic level, broadly corresponds to what is referred to as late Modern English. Although the English language may be considered to have remained almost intact at the phonological, morphological and syntactic levels over the two hundred years prior to the twentieth century, it does, however, experience a gradual but consistent development of a distinct scientific register during that time, with a specialised terminology and a distinctive genre of its own, the research article, following Boyle's (1661) ideas on the five compulsory characteristics it should present: 'brevity', 'lack of assertiveness', 'perspicuity', 'simplicity of form', and 'objectivity' (cf. Allen *et al.* 1994; Atkinson 1996; and Gotti 1996, 2001, 2003, 2005). The end of this period of development is also marked by linguistic change, with the early 1900s witnessing several arguments in favour of a new scientific register, such as that called for by Thomas Huxley at the 1897 'International Congress of Mathematics', resulting in the consolidation of a relatively standardised scientific register as we know it today.

## 3. General compilation criteria in the *Coruña Corpus*

Each sample included in the *Coruña Corpus* has been selected in such a way as to create a set of samples which mirror scientific writing (and each discipline) as faithfully as possible during the period, ensuring the representativeness of the corpus.

Representativeness is assured by means of two processes:

1. The selection of suitable specific texts as examples of genuine scientific writing comprising a series of requisites to be fulfilled in order to be considered for inclusion.
2. The conformation of a balanced selection of samples, including examples of different types of scientific writing being produced during the period, with the

aim of achieving, when considered as a whole, a balanced representation of the register during the Late Modern English period.

### *3.1. Criteria for inclusion of particular texts*

There are five main criteria that a text must satisfy to be considered a genuine manifestation of English scientific writing, and thus being eligible for inclusion in the *Coruña Corpus*.

First, only written, edited and published manifestations of scientific writing in prose are considered. Oral texts are excluded on the grounds that oral data is impossible to obtain for most of the period, although both transcriptions of lectures and scripts intended to be read aloud are eligible. Also excluded are texts in verse, since the inherent constraints in the language used in these texts imply a distorted or deliberately manipulated use of English, thus rendering such texts unrepresentative of the register.

Second, only texts written by native speakers are selected, since the use of English by non-native writers would not be representative of the English used in scientific writing during the period. Moreover, authors who completed all their training in English-speaking territories are prioritised on the assumption that these writers would be likely to present more genuine linguistic habits than those who lived and studied abroad.

In the same spirit, only texts written directly in English are selected, thus excluding translations, even where authors were the translators themselves, because interferences from the original language might have appeared in the translated text. This criterion is problematic, since a good proportion of the scientific production of the period was originally written in Latin, particularly at the start of the eighteenth century.

A further criterion is that only one work per author can be selected, thus avoiding jeopardising representativeness by over-representing the idiosyncrasies of particular authors. This limitation is applied at the corpus level rather than at the subcorpus level, so that only one work by any given author is selected for the whole of the *Coruña Corpus*.

Finally, first editions are preferred whenever possible, in order to avoid distorting the results on the diachronic axis by including samples from subsequent editions.

However, where first editions are not available, samples from editions published within a thirty-year timespan from the publication of the first edition are eligible, following Kytö *et al.*'s (2000: 92) assumption that thirty years is the minimum timespan in which language change can typically be observed.

### *3.2. Criteria to conform a balanced and representative set of samples*

In order to achieve the desired balance and representativeness across the whole set of samples, each sample has to be selected very carefully in relation with all other samples in the subcorpus. In order to do so, each eligible sample is classified according to a series of parameters. Alongside the discipline and the period of the text, these include genre, plus the sex and linguistic background of the author.

These parameters in the classification of each individual sample are compared with information drawn from a detailed consideration of the history of the discipline over this period, taking into account its particularities and characteristic uses. In this way we can achieve as faithful a representation of the reality of the register during the period as possible. Some relevant aspects of the development of early studies in Language and Linguistics are discussed in what follows.

## 4. LANGUAGE AND LINGUISTICS DURING THE EIGHTEENTH AND NINETEENTH CENTURIES

Although interest in language appears in the very earliest works of Philosophy, studies on Language and Linguistics would not emerge as a distinct discipline of study until the nineteenth century, when a growing interest in biological evolution and diversification brought about an inevitable curiosity in the evolution of the world's different languages and the ontological meaning and transcendence of language as such. For a very long time, the study of language had been restricted largely to Latin, the official language for both the church and academic activity, with little attention paid to vernaculars, which were considered mere tools for communication, or in the case of poetry as an endeavour related more to entertainment than culture (cf. Bailey 1985; Beal 2004, 2012; Crespo 2004).

In the seventeenth century, however, it became apparent that English was slowly but steadily gaining popularity as an object of intellectual curiosity thanks to the

coincidence of a number of factors. The expansion of the British Empire between the late-sixteenth and early-eighteenth centuries led to a rise in the status of the English language, which was now associated with power, prestige and wealth. Using good English now became a means of social advancement, and for those who wished to have a certain status in society it became important to speak and write correctly (cf. Beal 2004; Hickey 2010; Millward and Hayes 2012). In the eighteenth century, the possibility of rising economically (and, to some extent, socially) in the spheres of trade and commerce created a "linguistically insecure middle class" (cf. Beal 2008: 22–23), whose financial success appeared to depend largely on their mastery of the linguistic register of their culturally superior clients. On the other hand, the translation of the Bible and the progressive substitution of Latin by English in academic and other official contexts, which had in fact begun far earlier (cf. Taavitsainen and Pahta 1998), created the need for wider and more conscious instruction in the vernacular, with a consequent proliferation of grammars and manuals for correct usage and pronunciation.

The preoccupation of philologists with grammar became very apparent in the eighteenth century, as can be seen in specific works by Swift (1712), Stackhouse (1731), Johnson (1747), and Fisher (1753), all of which are included in our corpus. While some were particularly concerned with the correct use of spelling and syntax, this as a reflection of a more cultivated social status through writing, others were unhappy with a number of linguistic trends of the time, most of which were considered linguistic corruptions that needed to be corrected.[4] As a result, many of the English grammars in this period can be regarded as style manuals, in that they often included extended essays on the *status quo* of the English language, along with lists of 'corrupt' terms or expressions which they advised readers to avoid. On the other hand, a simultaneous interest in the etymology and internal organisation of the vernacular awakened in philologists a renewed interest in classical languages and in the way that these were approached, which itself led to several attempts to revise and improve Greek and Latin grammars and manuals (such as Sheridan 1714, or Squire 1741, also included in CETeL).

In the nineteenth century, the German linguist and philosopher Willhelm von Humbolt observed that human language was a rule-governed system, and as such deserved to be described (cf. Schmidt 1975; Di Cesare 1990). Already by the end of the

[4] In fact, both Swift (1712: 16) and Johnson (1747: 10) were rather pessimistic about language change.

1700s, language began to be treated as an object of study of natural sciences, and languages themselves were treated as living entities and thus classified into families according to their origins, their evolution, and their behaviour (cf. Campbell 2001). Heavily influenced during this period by Darwinism, the study of languages entailed the reconstruction of their origins back to Proto-Indoeuropean, culminating by the end of the century in the work of the Neogrammarians (cf. Robins 1978, 1997). At the same time, a growing interest in Asian languages and cultures – the result of a new scientific interest in the colonies (cf. De la Cruz Cabanillas 2001; Beal 2004) – led to the extension in the scope of ancient languages under study to those outside Europe, as well as in the increasing habit of working on more than one language at a time, a practice which opened the doors to modern Comparative Linguistics.

All the trends summarised above can be found in the samples included in CETeL, and some of these trends are directly related to specific challenges faced during the process of compilation. These difficulties will be described in the next section.

## 5. Criteria applied: Difficulties faced during the process of compilation

As described above, the selection of samples in all subcorpora of the *Coruña Corpus* is conducted in such a way as to make the set of samples representative of the disciplinary practices of the time, and the selection of samples in CETeL is no exception. However, in this case, the process has been particularly challenging, especially for the beginning of the period, as the result of several factors.

First, as already noted, the development of Linguistics as an individual discipline occurs comparatively late, and this affects the process of selection of samples particularly during the first decades of the eighteenth century. Looking at the opinions of authors here regarding the nature of their own works, as expressed in their abstracts and other introductory material, we can find labels such as ‘language’, ‘grammar’, or ‘etymology’, yet these are not always used in the same ways in which we might understand them today. To resolve this problem, we established the criterion that CETeL would be a corpus of texts on Language, rather than on Linguistics. Thus, CETeL goes beyond Linguistics as it would be considered nowadays, introducing several texts on the nature and philosophy of language, thus representing how scientific discourse on Language was considered at the time. Moreover, this also represents the

reality of scientific work during the period, as disciplines have become ever more specialised since the first decades of the eighteenth century (cf. Burke 2000: 132–137).

A second difficulty, although perhaps less important here than in other disciplines, was that Latin was still widely used in texts on Language and Linguistics well into the eighteenth century. This meant that a significant number of possible samples were ineligible, including not only samples written in Latin, but also samples in English which were translated from Latin. Identifying these translations was particularly difficult, because they were sometimes not advertised as such, particularly when the author was both the original writer and the translator. This made it necessary to conduct a comprehensive review of all the work of a given author in order to ascertain that a sample was indeed not a translation from a previous original in Latin or any other language.

Thirdly, some of the formats used during the period led to specific problems during the process of computerisation. For instance, dictionaries, with their organisation in entries which repeat the same grammatical structures, normally present little linguistic interest, whereas grammars of foreign languages, with a high number of examples in these languages, perhaps even in different alphabets, pose problems for transcription, as the rules of the *Coruña Corpus* qualify that the latter cannot be transcribed, and that the former, even if transcribed, have to be encoded in such a way that they do not count as words in the corpus (cf. Camiña and Lareo 2019: 22). This is also problematic in that whereas entire passages in a foreign language sometimes can easily be excluded, it is much more common to find foreign terms, endings, etc. inserted in the main text. In such cases, these are usually identified individually with editorial marks, but if they are so numerous as to impede transcription, the whole passage must be deleted, as it would not represent the real linguistic habits of the author. A further problem is the lack of a standard phonetic transcription for works on phonetics, and also in grammars for foreign languages where we find examples of how such words should be pronounced.

The final, and perhaps most notable, problem is the general lack of information about many authors at the beginning of the period. The principles of the *Coruña Corpus* state that it is preferable to select authors "about whom we could find basic biographical information and hence whose linguistic habits we could infer," (cf. Moskowich 2012: 48) and thus to be able to confirm that they are indeed native speakers. When this has

not been possible, samples were discarded, and this led to the rejection of a considerable number of otherwise valid samples from our initial inventory.

Further problems are related to the balance of the set of samples in the corpus. An important aspect of this balance is that it does not imply all the different categories being evenly represented throughout the period, which is, in any case, an impossible task with only two samples per decade. Rather, balance implies that the selection of samples should be representative of the reality of the discipline during the period, providing a good representation of both inter-disciplinary and intra-disciplinary differences across the different parameters.

This is best seen in the unequal distribution of genres over the period and across disciplines. For instance, as a result of the consideration of English as a means of social advancement, a large number of textbooks are included from both centuries.[5] This contrasts with other disciplines, such as Chemistry, in which textbooks were essential for the initial segment of the period, contributing to the dissemination of knowledge, but fell away in terms of importance during later stages.

However, sometimes not all categories are equally available, and this must also be taken into account when sorting the samples. For instance, particularly during the eighteenth century, grammars represent a very important proportion of all scientific production on Language. They reflect a widespread preoccupation with the correct use of language although, as noted above, some of these grammars also include a diachronic or stylistic perspective. Such works are featured in the corpus, but sometimes other content such as discussions on the correct use or the nature of language itself, both of which are also representative of the eighteenth century, are not as readily accessible as grammars. This leads to compilers having to choose samples from many valid grammars, whereas for other genres choosing among samples becomes impossible and it is sometimes necessary to include almost any valid sample. This in turn is complicated by the fact that sometimes genres are not easy to identify. A text might exhibit conflicting characteristics, being very broad and exhaustive in nature, and thus being potentially an example of either a treatise or a didactic work. In this sense, it could be classified as a textbook or as a manual but, in addition, it might have a question-and-answer format in a constructed dialogue form. In such cases, and if the

[5] It must be noted that the 'textbook' label includes a fuzzy textbook/handbook/manual category in this particular subcorpus.

author makes no reference to the genre itself, it is left to compilers to decide which genre is best represented in the sample. This is achieved by means of a close reading and a comparison of the texts with other, undoubted, texts to check the similarities between them. Such a comparison allows to assign the texts to a particular genre.

In order to faithfully represent the discipline of the period, several short texts have been included *in toto* even though they are shorter than the 10,000-word limit used across the *Coruña Corpus*, since they are characteristic of the production on linguistics in the period. However, special care has been taken in that the final number of words in any decade is roughly 20,000.

Finally, regarding the distribution of the samples according to the sex of their authors, it is worth mentioning that the majority of the samples of the *Coruña Corpus* are written by men, as was the case with science in general during the period. At that time, women faced considerable difficulties in accessing scientific knowledge and had to overcome a great many obstacles if they sought to become part of the social community of scientists. However, every subcorpus of the *Coruña Corpus* includes samples of texts written by women. Selecting female-authored texts has not always been easy, since publications by women lacked biographical information far more often than in the case of men, and women were also frequently obliged to write under pseudonyms or anonymously. Despite these difficulties, CETeL is among the subcorpora with the largest number of female authors.

## 6. DESCRIPTION OF CETeL

In this section, the beta version of CETeL will be described according to a number of parameters, namely: the distribution of the text samples over time, the topics and genres included in the overall set of samples, plus the sex and geographical origin and linguistic background of authors.[6]

CETeL contains a total of 44 samples, 24 from the eighteenth century and 20 from the nineteenth: the reason for this disparity lies in the inclusion of three, rather than two, samples in the following four decades: 1710s, 1720s, 1740s and 1780s. As shown in Table 1 below, most samples contain *c*.10,000 words, but in these four decades shorter

[6] It is important to note that the description provided here corresponds to the beta and not to the definite version of CETeL. Some classifications, particularly the word count of samples, may change during the process of revision, which is about to start.

texts were included. This was done partly due to the need to introduce some shorter texts *in toto* such as the 'Proposal for Correcting, Improving and Ascertaining the English Tongue' by Swift (1712) or Samuel Johnson's 'Plan of a Dictionary of the English Language' (1747), which were considered to be particularly representative of the period. In other cases, the quantity of text in a foreign language reduced the computable number of words in the selected text considerably, which was the case in 'The Rudiments of Grammar or the English-Saxon Tongue' by Elizabeth Elstob (1715). However, these issues do not affect the overall number of words, which is comparable in both centuries: 202,961 words in the eighteenth century and 203,062 in the nineteenth century.

| Date | Author | Title | Words |
|---|---|---|---|
| 1705 | Lane, Archibald | A key to the art of letters, or, English a learned language, full of art, elegancy and variety. Being an essay to enable both foreiners, and the English youth of either sex, to speak and write the English tongue well and learnedly, according to the exactest rules of grammar, after which they may attain to Latin, French, or any other forein language in a short time, with very little trouble to themselves or their teachers: with a preface shewing the necessity of a vernacular grammar. Dedicated to His Highness the Duke of Glocester. | 10,174 |
| 1706 | Johnson, Richard | Grammatical commentaries: being an apparatus to a new national grammar: by way of animadversion upon the falsities, obscurities, redundancies, and defects of Lilly's system now in use. | 9,908 |
| 1712 | Swift, Jonathan | A proposal for correcting, improving and ascertaining the English tongue, In a letter to the most honourable Robert Earl of Oxford and Mortimer, Lord High Treasurer of Great Britain. | 5,930 |
| 1714 | Sheridan, Thomas | An easy introduction of grammar in English for the Understanding of the Latin Tongue. Compil'd not only for the ease and encouragement of youth, but also for their moral improvement; having the syntaxis examples gathered from the choicest pieces of the best authors. To which is added a compendious method of variation and elegant disposition of Latin. | 7,777 |
| 1715 | Elstob, Elizabeth | The rudiments of grammar or the English-Saxon tongue, first given in English: With an apology for the study of Northern antiquities. Being very useful towards the understanding our ancient English poets, and other writers. | 6,839 |
| 1721 | Gildon, Charles | The Laws of Poetry, as laid down by the Duke of Buckinghamshire in his Essay on Poetry, by the Earl of Roscommon in his Essay on Translated Verse, and by Lord Lansdowne on Unnatural Flights in Poetry, Explain'd and Illustrated. | 6,161 |
| 1725 | Stevens, John | A new Spanish Grammar, more perfect than any hitherto published. All the errors of the former being corrected, and the rules for learning that language much improv'd. To which is added, a vocabulary of the most necessary words: Also a collection of phrases and dialogues adapted to familiar discourse. | 10,273 |
| 1728 | MacCurtin, Hugh | The Elements of the Irish Language, Grammatically Explained in English. In 14 chapters. | 5,140 |

Table 1: Samples included in CETeL and provisional word count in the beta version

| Date | Author | Title | Words |
|---|---|---|---|
| 1731 | Stackhouse, Thomas | Reflections on the Nature and Property of Languages in General, and on the Advantages, Defects and Manner of Improving the English Tongue in Particular. | 9,640 |
| 1737 | Greenwood, James | The Royal English Grammar: containing what is necessary to the knowledge of the English tongue. Laid down in a plain and familiar way. For the use of young gentlemen and ladys. | 10,014 |
| 1741 | Squire, Samuel | Two essays, the former a defense of the Ancient Greek Chronology; to which is annexed, a new chronological synopsis; the latter, an enquiry into the origin of the Greek Language. | 9,856 |
| 1747 | Johnson, Samuel | The plan of a dictionary of the English language: addressed to the Right Honourable Philip Dormer, Earl of Chesterfield; One of His Majesty's Principal Secretaries of State. | 6,909 |
| 1748 | Martin, Benjamin | Institutions of Language; Containing, a physico-grammatical Essay on the propriety and rationale of the English tongue. Deduced from A general idea of the nature and necessity of speech for human society; A particular view of the genius and usage of the original mother tongues, the Hebrew, Greek, Latin, and Teutonic; with their respective idioms, the Italian, French, Spanish, Saxon, and German, so far as they have relation to the English tongue, and have contributed to its composition. | 10,138 |
| 1751 | Harris, James | Hermes: Or, a philosophical inquiry concerning language and universal grammar. | 11,350 |
| 1753 | Fisher, Anne | A new grammar, with exercises of bad English: or, An easy guide to speaking and writing the English language properly and correctly. | 9,841 |
| 1762 | Priestley, Joseph | A Course of Lectures on the Theory of Language and Universal Grammar. | 8,855 |
| 1765 | Elphinston, James | The Principles of the English Language Digested, or, English Grammar Reduced to Analogy. | 11,604 |
| 1771 | Fenning, Daniel | A New Grammar of the English Language; or, an easy introduction to the art of speaking and writing English with propriety and correctness: The whole laid down in the most plain and familiar manner, and calculated for the use, not only of schools, but of private gentlemen. | 8,617 |
| 1776 | Campbell, George | The Philosophy of Rhetoric. | 9,082 |
| 1784 | Nares, Robert | Elements of orthoepy: containing a distinct view of the whole analogy of the English Language; so far as it relates to pronunciation, accent, and quantity. | 10,058 |
| 1784 | Webster, Noah | A Grammatical Institute of the English Language, comprising, an easy, concise, and systematic method of education, designed for the use of English schools in America. In three parts. | 10,040 |
| 1786 | Jones, William | The Third Anniversary Discourse, on the Hindus. Delivered 2 February, 1786. By The President. | 4,687 |
| 1797 | Tytler, Alexander Fraser | Essay on the Principles of Translation. | 10,068 |
| 1798 | Fenn, Eleanor | The mother's grammar. Being a continuation of the child's grammar. With lessons for parsing. And a few already done as examples. | 9,350 |
| 1810 | Adams, John Quincy | Lectures on rhetoric and oratory: delivered to the classes of senior and junior sophisters in Harvard University. | 11,913 |

Table 1 (continuation)

| Date | Author | Title | Words |
|---|---|---|---|
| 1810 | Smart, B. H. | A practical grammar of English pronunciation: on plain and recognised principles, calculated to assist in removing every objectionable peculiarity of utterance, arising rather from foreign, provincial or vulgar habits; or from a defective use of the organs of speech; and furnishing, to pupils of all ages, the means of systematically acquiring that nervous and graceful articulation, which is the basis of a superior delivery: together with directions to persons who stammer in their speech, comprehending some new Ideas relative to English prosody. | 9,611 |
| 1815 | Richardson, Charles | Illustrations of English philology. | 8,425 |
| 1819 | Cobbett, William | A grammar of the English language: in a series of letters. Intended for the Use of Schools and of Young Persons in general; but, more especially for the Use of Soldiers, Sailors, Apprentices and Plough-boys. | 12,713 |
| 1825 | Cardell, William S. | Essay on language: as connected with the faculties of the mind, and as applied to things in nature and art. | 15,040 |
| 1830 | Booth, David | An analytical dictionary of the English language; in which the words are explained in the order of their natural affinity, independent of alphabetical arrangement; and the signification of each is traced from its etymology, the present meaning being accounted for when it differs from its former acceptation: the whole exhibiting, in one continued narrative, the origin, history, and modern usage of the existing vocabulary of the English tongue: to which are added, an introduction, containing a new grammar of the language, and an alphabetical index, for the ease of consultation. | 11,026 |
| 1836 | Allen, Alexander | An etymological analysis of Latin Verbs. For the use of schools and colleges. | 10,128 |
| 1836 | Bosworth, Joseph | The origin of the Germanic and Scandinavian languages, and nations: with a sketch of their literature, and short chronological specimens of the Anglo-Saxon, Friesic, Flemish, Dutch, the German from the Mœso-goths to the present time, the Icelandic, Danish, Norwegian and Swedish: tracing the progress of these languages, and their connexion with the Anglo-Saxon and the present English. With a map of European Languages. | 10,601 |
| 1841 | Latham, Robert Gordon | Elements of the English Language for the use of Ladies' Schools. | 10,061 |
| 1845 | Ellis, Alexander John | The Alphabet of Nature or contributions towards a more accurate analysis and symbolization of spoken sounds; with some account of the principal phonetical alphabets hitherto proposed. | 10,237 |
| 1852 | Rawlinson, Sir Henry Creswicke | Outline of the History of Assyria, as collected from the inscriptions discovered by Austin Henry Layard, Esq. In the Ruins of Nineveh. Printed from the Journal of the Royal Asiatic Society. | 11,139 |
| 1854 | Baker, Anne Elizabeth | Glossary of Northamptonshire Words and phrases, with examples of their colloquial use, and illustrations from various authors to which are added, the customs of the county. | 10,069 |
| 1867 | Whitney, William Wight | Language and the Study of Language: Twelve Lectures on the Principles of Linguistic Science. | 10,196 |
| 1870 | Steere, Edward | A Handbook of the Swahili Language as Spoken at Zanzibar. | 10,066 |
| 1871 | Earle, John | The Philology of the English Tongue. | 10,639 |

Table 1 (continuation)

| Date | Author | Title | Words |
|---|---|---|---|
| 1879 | Findlater, Andrew | Language (Chambers's Elementary Science Manuals) | 10,812 |
| 1880 | Bain, Alexander | Higher English Grammar. | 10,109 |
| 1886 | Bell, Alexander Melville | Essays and postscripts on elocution. | 10,108 |
| 1891 | Dickson White, Andrew | New Chapters in the Warfare of Science, XI. From Babel to Comparative Philology | 10,230 |
| 1892 | Sweet, Henry | A Short Historical English Grammar. | 10,135 |
| | | TOTAL | **406,023** |

Table 1 (continuation)

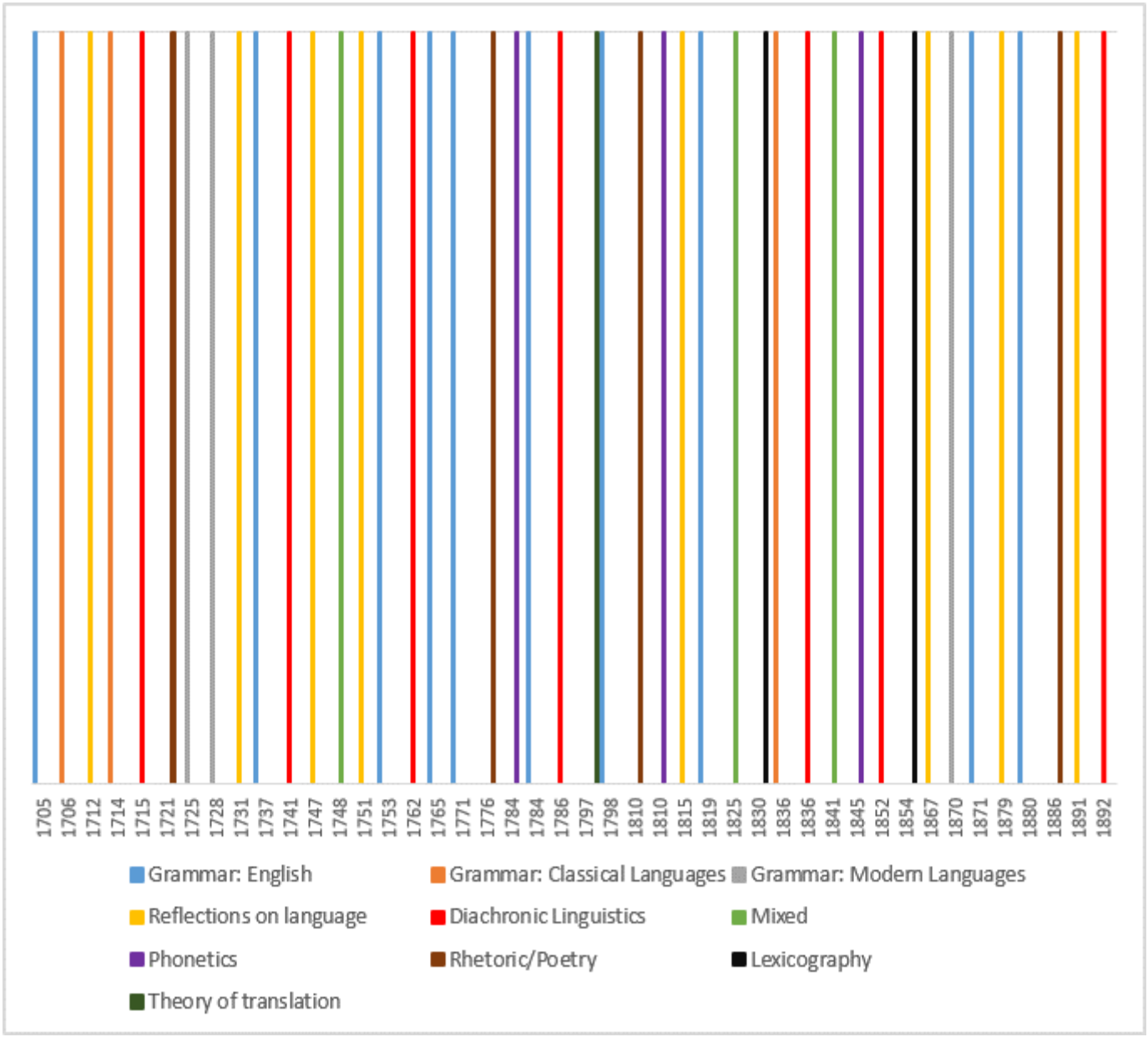


Figure 2: Distribution of topics in the samples of CETeL over time

The distribution of topics in the samples, as shown in Figure 2 above, allows us to see the evolution of the discipline over time. For instance, English grammars, shown in light blue, are found throughout the whole period, whereas grammars of classical languages (orange) are concentrated in the early 1700s, with only another example in the 1830s. As mentioned in Section 4 above, some of those grammars also contain diachronic explanations, and they are labelled as 'mixed topic' (light green). There is also a notable number of grammars of modern languages (grey) which also appear in both centuries.

Other subjects, such as works on Phonetics (purple), Rhetoric (brown) and Lexicography (black), emerge somewhat later in the period. The first of these, of which we have three samples, only begins in the 1780s, just a decade later than works on Rhetoric (1770s, as the first sample in the 'rhetoric and poetry' group, in the 1720s, deals with Poetry). The two lexicographical works – a dictionary and a phrasebook of localisms – are both from the nineteenth century. A work dealing with Theory of translation (dark green) appears at the very end of the eighteenth century.

Red bars represent works on what is nowadays considered Diachronic Linguistics. As can be seen in Figure 2, this was a topic which received attention throughout the period, although the treatment of the topic changed considerably from the early eighteenth century, with an interest in pureness against corruption, to the end of the nineteenth, with efforts towards reconstruction using the comparative method. On the other hand, yellow bars represent what has provisionally been labelled 'reflections on Language', and once more these are present throughout the whole period. These are texts which deal with concepts of Language and Linguistics, and thus samples are drawn from texts that discuss the correction of English (or, rather, denounce its corruption), as well as from theoretical works on the nature of language, its origins, or the philosophical matter behind them, which appear to be the first works on the discipline of Linguistics as we understand it today.

Regarding the genres of the samples, a preliminary classification (cf. Figure 3, below) shows that the most frequently represented genre – twelve examples in total – is that of textbooks, followed by treatises and essays, this reflecting the didactic and, up to a point, philosophical nature of the works written on Language during the period. Likewise, there is a relatively high number of lectures (five), several of which correspond to speeches written to be delivered at meetings of the various societies

established for the study of Language in both the eighteenth and the nineteenth centuries. The number of letters (four) reflects the well-known use of this genre in scientific discourse during the period, particularly in the eighteenth century, and indeed the subcorpus keeps with historical use here in that the last included letter dates from 1819.

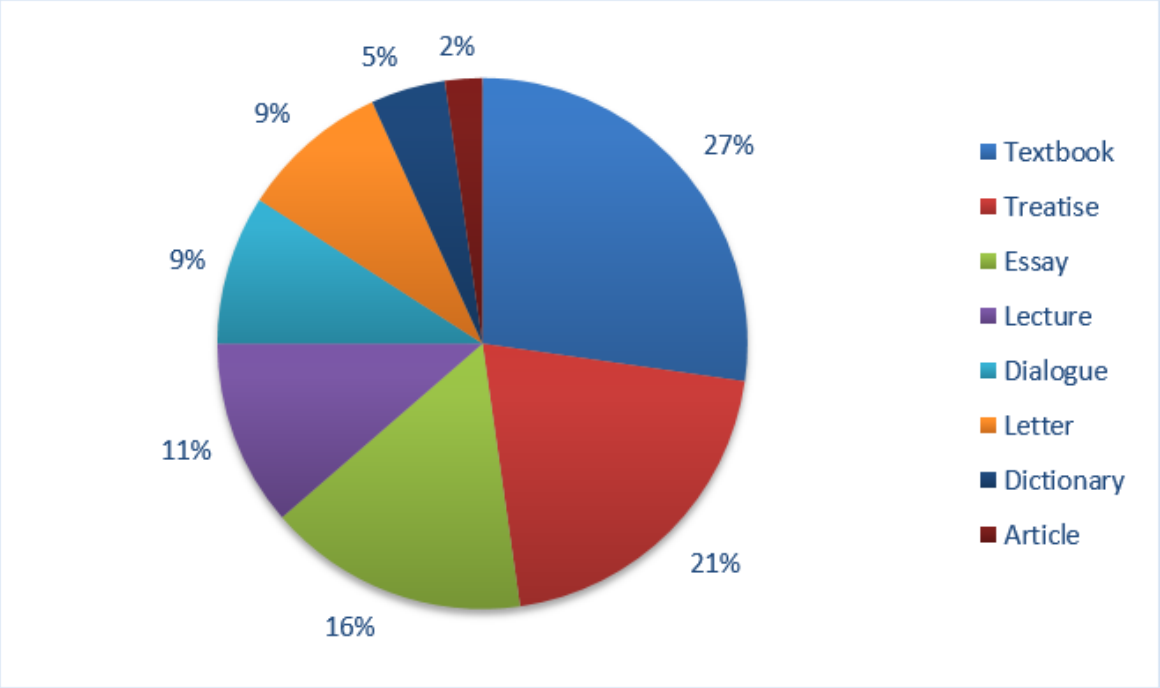


Figure 3: Genres in the samples in CETeL

Among the least frequent genre are two dictionaries, a very specific genre which appears to be relatively frequent in the discipline compared to others, as there are only three other examples of dictionaries (for Astronomy, History, and Chemistry) in the rest of the *Coruña Corpus* to date. Finally, there is only one article, dating from 1891, this reflecting the comparatively late emergence of this genre in the discipline.

Figure 3 also shows four dialogues, which merit special attention. These are different from other dialogues included in other subcorpora of the *Coruña Corpus* (cf. CETA and CEPhiT), in that rather than presenting a conversation between two or more characters, they comprise series of questions and answers, similar to catechisms, albeit of a non-religious kind. All four dialogues follow this format, which raises the question as to whether they should be considered dialogues or, rather, it might be necessary to create a new category for this putative genre. However, since they contain similar

structures to other dialogues in the *Coruña Corpus*, and there are no contrasting samples (either dialogues with this format in other disciplines, or dialogues with any other format), it was decided to classify them as part of the category 'dialogue'.

Regarding the sex of the authors, Figure 4 shows that only four of the 44 samples included in CETeL were written by women. This represents 9.09% of the total samples, which seems representative of the discipline in the period under study. This represents a higher proportion than in other subcorpora (cf. CETA 4.76%, CEPhiT 7.5%, CECheT 7.31%), but a far lower proportion than in CELiST and CHET, both with 20% of female-authored samples. These percentages are in keeping with the aim of representativeness, both in the whole corpus and in each discipline, for Life Sciences and History were among the disciplines which were most open to female practitioners.

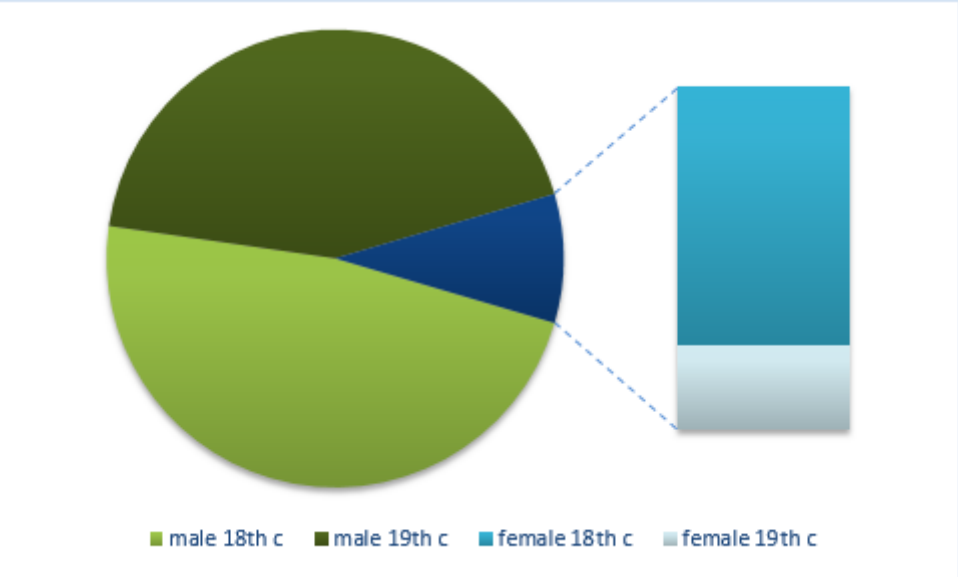


Figure 4: Samples per sex in CETeL

Finally, in terms of the geographical origin and linguistic background of authors, Figure 5 below shows that most samples were written by English authors, followed by Scottish, North American and Irish ones. The four samples marked 'other' include authors for whom little or no information has been found, or who were educated in more than one place, making it very challenging for compilers to ascertain where they might have acquired their linguistic habits.

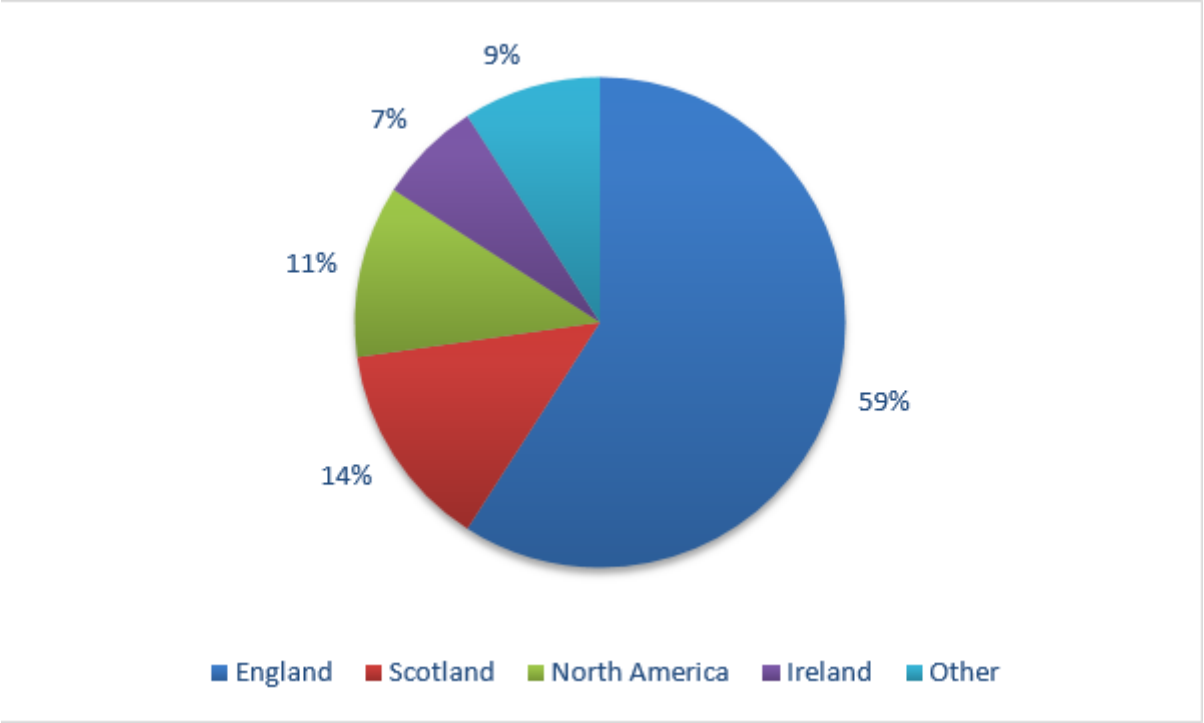


Figure 5: Geographical origin per sample in CETeL

## 7. Concluding remarks

This paper has presented a new subcorpus of the *Coruña Corpus*, namely CETeL, focusing on its main characteristics regarding the timespan, topics, and genres of its text samples, and the sex and linguistic background of the authors. It has also pointed out the main drawbacks and challenges faced in the process of compilation. Once published, CETeL, the first corpus of its kind on Language and Linguistics during the eighteenth and nineteenth centuries, is expected to be a reliable source of linguistic data for research on the evolution of the English linguistic subregister throughout the Late Modern English period, as well as a valuable illustration of historical scientific writing. With the process of computerisation being now complete, a process of revision is underway, in which each of the samples will be manually revised three times by different reviewers in order to guarantee the most faithful representation of the original. CETeL is scheduled to be completed over the 2020–2022 period, although final beta versions for testing will be made available sooner.

# REFERENCES


Allen, Bryce, Jian Qin and Frederik Wilfrid Lancaster. 1994. Persuasive communities: A longitudinal analysis of references in the philosophical transactions of the Royal Society, 1665–1990. *Social Studies of Science* 24/2: 279–310.

Atkinson, Dwight. 1996. The philosophical transactions of the Royal Society of London, 1675–1975: A sociohistorical discourse analysis. *Language in Society* 25/3: 333–371.

Bailey, Richard W. 1985. The conquests of English. In Sidney Greenbaum ed. *The English Language Today*. Oxford: Pergamon Institute of English, 9–19.

Beal, Joan. 2004. *English in Modern Times*. London: Arnold.

Beal, Joan. 2008. Shamed by your English? The market value of a 'good' pronunciation. In Joan Beal, Carmela Nocera and Massimo Sturiale eds. *Perspectives on Prescriptivism*. Bern: Peter Lang, 21–40.

Beal, Joan. 2012. Late Modern English in its historical context. In Isabel Moskowich and Begoña Crespo eds. *Astronomy 'Playne and Simple.' The Writing of Science between 1700 and 1900*. Amsterdam: John Benjamins, 1–14.

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

Biber, Douglas and Susan Conrad. 2009. *Register, Genre, and Style.* Cambridge: Cambridge University Press.

Boyle, Robert. 1661 (1965). Proemial essay. In Thomas Birch ed. *The Works of Robert Boyle*. Vol. I. Hildesheim: Georg Olms, 192–204.

Burke, Peter. 2000. *Historia Social del Conocimiento: De Gutemberg a Diderot.* Vol. I Barcelona: Paidos Ibérica.

Camiña, Gonzalo and Inés Lareo. 2019. Editorial policy in CHET. In Isabel Moskowich, Estafanía Sánchez-Barreiro, Inés Lareo and Paula Lojo-Sandino comps eds. *Corpus of History English Texts (CHET)*. A Coruña: Repositorio Universidade da Coruña. https://ruc.udc.es/dspace/handle/2183/21849 (29 September, 2019)

Campbell, Lyle. 2001. The history of linguistics. In Mark Aronoff and Janie Rees-Miller eds. *The Handbook of Linguistics*. Oxford: Blackwell, 81–104.

Claridge, Claudia, Josef Schmied and Rainer Siemund. 1999. The Lampeter Corpus of Early Modern English tracts. In Knut Hofland, Anne Lindebjerg and Jørn Thunestvedt eds. *ICAME Collection of English Language Corpora* (CD-ROM). Norway: The HIT Centre, University of Bergen.

Crespo, Begoña. 2004. The scientific register in the history of English: A corpus-based study. *Studia Neophilologica* 76/2: 125–139.

De la Cruz Cabanillas, Isabel. 2001. Lexicografía y semántica del inglés moderno. In Isabel de la Cruz Cabanillas and Francisco Javier Martín Arista eds. *Lingüística Histórica Inglesa*. Barcelona: Ariel, 699–727.

Di Cesare, Donatella. 1990. The philosophical and anthropological place of Wilhelm von Humboldt's linguistic typology: Linguistic comparison as a means to compare the different processes of human thought. In Tullio De Mauro and Lia Formigari eds. *Leibniz, Humboldt, and the Origins of Comparativism*. Amsterdam: John Benjamins, 157–179.

Gotti, Maurizio. 1996. *Robert Boyle and the Language of Science*. Milano: Guerini Scientifica.

Gotti, Maurizio. 2001. The experimental essay in Early Modern English. *European Journal of English Studies* 5/2: 221–239.

Gotti, Maurizio. 2003. *Specialized Discourse: Linguistic Features and Changing Conventions*. Bern: Peter Lang.
Gotti, Maurizio. 2005. *Investigating Specialized Discourse*. Bern: Peter Lang.
Gray, Bethany. 2011. *Exploring Academic Writing through Corpus Linguistics: When Discipline Tells only Part of the Story*. Flagstaff, AZ: Northern Arizona University (Unpublished PhD dissertation).
Hickey, Raymond. 2010. Attitudes and concerns in eighteenth-century English. In Raymond Hickey ed. *Eighteenth-Century English*. Cambridge: Cambridge University Press, 1–19.
Kytö, Merja, Juhani Rudanko and Erik Smitterberg. 2000. Building a bridge between the present and the past: A corpus of 19th-century English. *ICAME Journal* 24: 85–97.
Millward, Celia M. and Mary Hayes. 2012. *A Biography of the English Language.* Boston: Wadsworth, Cengage Learning.
Moskowich, Isabel. 2012. CETA as a tool for the study of modern astronomy in English. In Isabel Moskowich and Begoña Crespo eds. *Astronomy 'Playne and Simple.' The Writing of Science between 1700 and 1900*. Amsterdam: John Benjamins, 35–56.
Moskowich, Isabel and Begoña Crespo eds. 2012. *Astronomy 'Playne and Simple.' The Writing of Science between 1700 and 1900*. Amsterdam: John Benjamins.
Moskowich, Isabel, Gonzalo Camiña-Rioboo, Inés Lareo and Begoña Crespo eds. 2016. *The Conditioned and the Unconditioned: Late Modern English Texts on Philosophy*. Amsterdam: John Benjamins.
Moskowich, Isabel, Begoña Crespo, Luis Puente-Castelo and Leida Maria Monaco eds. 2019. *Writing History in Late Modern English: Explorations of the Coruña Corpus.* Amsterdam: John Benjamins.
Robins, Robert H. 1978. The Neogrammarians and their nineteenth-century predecessors. *Transactions of the Philological Society* 76/1: 1–16.
Robins, Robert H. 1997. *A Short History of Linguistics*. London: Routledge.
Schmidt, Siegfried. 1975. German philosophy of language in the late 19th century. In Herman Parret ed. *History of Linguistic Thought and Contemporary Linguistics*. Berlin: de Gruyter, 658–684.
Taavitsainen, Irma and Päivi Pahta. 1998. Vernacularisation of medical writing in English: A corpus-based study of scholasticism. *Early Science and Medicine* 3/2: 157–185.

*Corresponding author*
Leida Maria Monaco
University of Oviedo
Department of English, French and German
Calle Amparo Pedregal, 5
33011 Oviedo
Spain
e-mail: lmonaco@uniovi.es

received: August 2018
accepted: October 2019

RiCL Research in Corpus Linguistics

# Koder – A multi-register corpus for investigating register variation in contemporary German

Andressa Costa
PUC São Paulo / Brazil

**Abstract** – This paper introduces the design decisions in building the Koder corpus, a multi-register-corpus of contemporary German. The purpose of this corpus is to serve as a basis for the investigation into the use of German across registers. In order to construct a representative corpus, the essential considerations are: the type and number of registers to include, the number of texts in each register and minimal text length. The paper describes which aspects were central in determining these issues as well the corpus composition and the necessary text processing.

**Keywords** – corpus design; Koder; register; German

## 1. INTRODUCTION

The availability of corpora facilitates the investigation of language use considerably. At present, there are various German corpora available to the academic community. In spite of this, building a corpus is sometimes still necessary because they are not completely suitable for answering some research questions. This paper describes the design decisions and composition of Koder (*Korpus deutscher Register*). The purpose of this corpus is to serve as a basis for empirical investigations of the German language through different registers. Most studies look at linguistic phenomena only in one register or they investigate only spoken, written documents or documents from the Internet. Therefore, available corpora from German are neither diversified in terms of mode nor cover a wide range of registers. Nevertheless, materials from available corpora from the Institute for the German Language (IDS), *Dortmunder-Chat-Korpus* (Beißwenger 2013) and *German Political Speeches* (Barbaresi 2012) were integrated in this corpus.

*Research in Corpus Linguistics* 7: 69–83 (2019). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.07.04

The necessity for building this corpus comes from the intention to investigate some linguistic phenomena across different registers because, as Biber and Conrad (2009: 6–7) observe, the use of linguistic features is influenced by the register in which they are being used. Register, as used in this study, refers to "a variety associated with a particular situation of use (including particular communicative purpose)" (Biber and Conrad 2009: 6).

A central aspect to consider when building a corpus is representativeness. It involves determining the corpus size, that is, the number and types of texts to be included in the corpus, the number of words per text and the total number of words in the corpus as well as the types of registers, in the case of a multi-register corpus (Berber Sardinha 2004: 24–25). The decisions made on these aspects will depend on the goals of the analysis. However, more important considerations than corpus size are a definition of the target population and choices concerning the method of sampling (cf. Biber 1993a). In fact, Biber (1993a: 243) defines representativeness as "the extent to which a sample includes the full range of variability in a population."

As Biber (1993b: 219) notes, two kinds of error must be minimised to achieve a representative corpus: 'random error' and 'bias error'. A random error occurs when a sample is not large enough to accurately estimate the right population; a bias error is when the selection of a sample is systematically different from the population. Thus, one important consideration relates to how to sample language in a corpus to study general language. On this issue, Biber (1993b: 220) argues that "analyses must be based on a diversified corpus representing a wide range of registers to be appropriately generalised to the language as a whole." He justifies this view with the assumption that there is no adequate overall linguistic characterisation of an entire language; instead, there are marked linguistic differences across registers. In order to select the registers that adequately represent a language, it is necessary to consider the users of that language. Regarding corpus size, as Sinclair (2005) states, there is no maximum size. The author considers two main factors in establishing the minimum size of a corpus: "1. the kind of query that is anticipated from users; and 2. the methodology they [the researchers] used to study the data." To analyse linguistic variation using a corpus-based approach, Biber (1990) provides an empirical investigation with the following methodological issues regarding corpus construction:

1) How long texts should be to reliably represent the distribution of linguistic features in particular text categories;
2) How many texts within each text category are required to reliably represent the linguistic characteristics of that category and related questions concerning the validity of register categories;
3) How many texts are needed in a corpus to accurately identify the salient parameters of variation among texts;
4) How much of a cross-section is required to identify and analyse the salient parameters of variation among texts.

In his investigation, Biber (1990: 261–268) analyses and compares samples of different sizes using statistical techniques. The results indicate the following:

– There is a high level of stability for the analysed linguistic features in 1,000-word sub-samples of texts so that 2,000-word and 5,000-word texts in the standard corpora are reliable representatives of their text categories;
– 10-text sub-samples accurately represent the linguistic characteristics of register categories, including both the central tendency and the range of variation;
– A factor-analysis with a corpus of 120 texts and another with a corpus of 240 texts containing the full range of registers included in the original corpus (23 registers) reasonably well represents the underlying parameters of variation that were found in the initial factor analysis with a corpus of 481 texts;
– A corpus of 169 texts, with fewer registers than the other two samples, provides a poorer representation of the underlying parameters found in the original corpus.

Biber (1990: 269) showed in this study that "the underlying parameters of text-based linguistic variation […] can be replicated in a relatively small corpus if that corpus represents the full range of variation." Berber Sardinha (2004) applied the methodological procedures suggested by Biber (1990, 1993a) and proposed the minimum number of approximately 5,500,000 words for a general corpus of English and nearly 91,000 for a specific corpus. For their investigation of register variation in Brazilian Portuguese, Berber Sardinha *et al*. (2014) built a multi-register-corpus of 48 registers with 20 texts per register and texts with at least 400 words. The decisions made in designing Koder were based on the works of Biber (1990, 1993a, 1993b) and Berber Sardinha (2004) for determining the number of registers, as well the number of texts

within the register and minimal text length. The register selection was based on the typology developed in several chapters in Brinker *et al*. (2000) and Eroms (2008).

## 2. KODER (KORPUS DEUTSCHER REGISTER)

### *2.1. Register selection*

The first step for register selection was to identify which registers are productive and represent the range of situational variation in contemporary German. This was not an easy task because there is no source where this information can be found. However, the typology of fields of communication presented in several chapters in Brinker *et al.* (2000) and Eroms (2008) for written and spoken texts served as a starting point. Brinker *et al.* (2000: XXVI) define fields of communication as an 'ensemble' of text types that are situationally and socially defined. This definition is to some extent similar to the definition of register adopted by Biber and Conrad (2009: 5), who consider register a category of texts with shared situational characteristics, whereas dialects are defined as a category of texts with shared social characteristics. The term 'field of communication' is not yet established, as Adamzik shows (2016: 126). Nevertheless, it was useful information to begin the selection of the register for this project.

The list of fields of communication proposed in Brinker *et al*. (2000), though comprehensive, has been considered provisional and unsystematic because an adequate typology for German texts has yet to be established. Moreover, it comprises only fields for written communication, excluding computer-mediated communication. Documents from the Internet and other registers like movies and non-fiction, which are not on the proposed typology, were added to this project. The selection of internet registers was based on Beißwenger and Lemnitzer (2013) and Berber Sardinha (2014), and the selection of movies on Veirano Pinto (2013). The second step was to determine the amount of registers. Because this corpus is currently being used as a basis for investigations about the general use of German and about individual linguistic features, the decision was made to include the complete range of registers described by the consulted literature.

Certain registers were selected from sources other than the consulted literature. This is the case for the registers under the label 'others' and the label 'oral communication' which comprises two categories from the *Database for Spoken*

*German* (DGD): *Forschungs- und Lehrkorpus* (FOLK) and *Gesprochene Wissenschaftssprache* (GWISS). Other registers from this database included in the corpus are conversations, oral exams and academic lectures. Material from Facebook and Twitter was collected from public accounts rather than from private users. The transcripts from TED talks subtitles were edited manually because they are automatically generated and contain many errors.

The registers included in this corpus represent a broad range of communicative situations in contemporary German, to which German speakers are currently exposed. It is not only diversified in terms of registers but also in terms of mode: the collection comprises written and spoken texts, as well as texts produced in a digital environment.

### *2.2. Text collection and corpus size*

After the selection of registers, text size and the number of texts had to be determined. For this purpose, two aspects were taken into consideration (Biber 1990: 258):

1. How many texts within each text category are required in order to represent the linguistic characteristics of that category reliably and the validity of register categories;
2. How long texts should be in order to reliably represent the distribution of linguistic features in a particular text category.

The first decision made was to build a balanced corpus in which all registers have the same number of texts. In order to determine how many texts each register should contain other studies using multi-register corpora served as orientation (Biber 1988; Biber *et al.* 2006; Xiao 2009; Berber Sardinha *et al.* 2014). Most of these studies did not use a balanced corpus except for Berber Sardinha *et al.* (2014), who used a corpus composed of 20 texts per register and included texts with at least 400 running words. The decision made for Koder was to collect 50 texts of at least 400 words for each register in order to build a corpus as large as possible in a limited time.

Nevertheless, some registers have more than 50 texts, whereas others have fewer than that. The reason why some registers, such as TED talks, detective series, and academic lectures, have fewer than 50 texts lies in the difficulty to find enough available material. Other registers have fewer than the minimum number of words established as part of the corpus design criteria (at least 400 words per text). Because

several texts from news, recipes, readers' letters to the editor, job advertisements, and song lyrics have fewer than 400 words, more texts were added to reach the minimum word length. Except for job advertisements and news, the following criteria were settled for the addition more texts:

- Recipes: 50 dishes were selected and two or more different recipes for each dish were collected;
- Readers' letter to the editor: 50 editions from magazines and newspapers were selected and all readers' letters to the editor were collected;
- Song lyrics: 50 singers or bands were selected and three songs from each singer or band were collected.
- Some internet registers have the same problem regarding text length. The decision in this case was the following:
  - Twitter: sets of tweets from about 50 different hashtags;
  - Facebook comments: sets of comments from 50 different posts;
  - YouTube comments: sets of comments from 50 different videos;
  - Reader commentary: sets of comments from about 50 different articles;
  - Wikipedia user talk: sets of comments from the editors of about 50 different Wikipedia articles.

The first purpose of this corpus is to serve as a basis in an investigation on register variation through the multi-dimensional approach (Biber 1988) in which a factor analysis is conducted in order to identify which linguistic features significantly co-occur in the specific registers. A pilot study undertaken to test the data revealed that the sample size (Kaiser-Meyer-Olkin measure of sampling adequacy = .84) is very good for conducting a factor analysis. Thus, Bartlett's Test of Sphericity (< .0001) shows that the correlation between the variables in the data is significantly different from 0, which means that they are suitable for a factor analysis (Loewen and Gonulal 2015: 187–188).

### *2.3. Text selection and compilation*

The decision about which texts to compile depended upon the availability of the materials. This criterion includes both available corpora and the permission to use material found on the Internet. Firstly, a list of text types was made on the basis of the

literature.[1] Subsequently, a search for available corpora was made and the texts of these corpora were collected. Afterwards, the availability of other text types to be collected without any legal restrictions was checked. Most of the texts were collected from the Internet and some of them were scanned.

Documents were collected from available corpora as follows: conversation, institutional communications and interviews were collected from FOLK; oral exams and academic lectures were collected from GWISS; Wikipedia user talk were compiled from *Deutsches Referenzkorpus* (DeReKo); professional chats from the *Dortmunder-Chat-Korpus* (Beißwenger 2013) and *German Political Speeches* is a corpus developed by Barbaresi (2012). The FOLK and GWISS corpora as well DeReKo are provided by *Institute for the German Language* (IDS). Material of the majority of registers was completely compiled from the Internet except for material of editorial and readers' letters to the editor which were partially scanned and partially compiled from the Internet.

The compilation of texts from the Internet involved the following criteria: a) a survey was undertaken in order to list newspapers, magazines, publishers, institutions, companies, websites about recipes, blogs, etc. from Germany and with the domain *.de*; b) the author of documents, such as academic texts and articles from newspapers and magazines as well as fictional literature, had to be German. When the author's origin could not be checked, the text was discarded. However, it was difficult to apply these criteria to Tweets and commentaries from Facebook and YouTube. In this case, the material was still collected. It is important to note here that the data from these three registers was gathered from public profiles. No data from personal profiles was collected.

The register academic and scientific institutions comprises two sub-registers: academic texts and popular science. Academic texts contain three different text types but only doctorate theses are split into groups: one group is composed of documents from Human and Social Sciences, the other group of documents from Natural, Engineering and Biological Sciences. In the collection, there are exclusively academic articles from Human and Social Sciences because it is difficult to obtain academic articles from Natural, Engineering and Biological Sciences written in German. It seems to be a tendency in such disciplines to write articles in English rather than in German. In

[1] See Tables 1 and 2 below.

contrast, popular science from Natural, Engineering and Biological Sciences articles which are written in German could be easily found. Academic textbooks are extracts which could only be found by one publisher. There is not much material available on the Internet: 38 texts are from Human and Social Sciences and 12 texts from Natural, Engineering and Biological Sciences. Similar to academic textbooks, the texts from fictional literature and non-fiction are extracts compiled from the websites of different German publishers.

Documents from media registers were selected from different national newspapers and magazines except for spoken news and news. Spoken news was collected from a German broadcaster which provides the transcriptions of the news on the website. The category news, which comprises short news, was collected from regional newspapers from different regions in Germany.

For the compilation of song lyrics, the following criteria were adopted: a) 50 singers and bands were selected from hit lists; b) research about the artists was made in order to select three songs by each artist which were composed between 1990 and 2018. The music genres are diverse: pop, rock, hip hop and rap.

The selection of movies and series occurred in two phases. Firstly, a list of German movies and series was made through a search on the web; secondly, a search for subtitles of the listed movies and series was conducted. The final selection contains the material which could be found. The variety of German series and films could not be successfully represented in this corpus because of a lack of available subtitles.

After the selection of texts described in the forerunning, the corpus content is summarised in Table 1 and Table 2. The documents are grouped into two categories: written (Table 1) and spoken registers (Table 2). The registers are categorised according to their fields of communication. Some registers have various text types, which are also identified in the tables. Moreover, the texts are classified according to features as dialogue/monologue, scripted/non-scripted, public/private, etc.

| Mode | Fields of Communication | Register | Sub-registers | Setting | Specific text-type included in corpus | Texts | Number of words |
|---|---|---|---|---|---|---|---|
| WRITTEN | University and Scientific fields (Heinemann 2000a) | Academic texts | Academic articles (Human and Social Sciences) | Public | Article | 50 | 287,052 |
| | | | Academic textbook (Human and Social Sciences) | Public | Textbook | 32 | 164,517 |
| | | | Academic textbook (Natural Sciences and Engineering) | Public | | 7 | 29,144 |
| | | | Academic textbook (Biological Sciences) | Public | | 11 | 55,523 |
| | | | Doctoral thesis (Human and Social Sciences) | Public | Doctoral thesis | 50 | 3,298,682 |
| | | | Doctoral thesis (Natural Sciences and Engineering) | Public | | 39 | 1,339,736 |
| | | | Doctoral thesis (Biological Sciences) | Public | | 11 | 332,142 |
| | | Popular science | Specialised/Technical texts (Natural Sciences and Engineering) | Public | Article | 27 | 51,010 |
| | | | Specialised/Technical texts (Biological Sciences) | Public | | 23 | 47,862 |
| | Medicine and Health (Wiese 2000) | Package insert | | Public | Manual | 50 | 91,608 |
| | Political Institutions (Klein 2000) | Plenary minutes | | Public | Minute | 50 | 551,938 |

Table 1: Written section of Koder (*Korpus deutscher Register*)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| WRITTEN | Documents from the web (Beißwenger and Lemnitzer 2013; Berber Sardinha 2014) | Blog | Professional chat | Public | | 50 | 45,087 |
| | | Website | | Public | | 50 | 36,003 |
| | | Wikipedia article | | Public | | 50 | 253,394 |
| | | Wikipedia user talk | | Interactive /Public | | 50 | 77,905 |
| | | Chat | | Interactive /Public | | 50 | 100,395 |
| | | Facebook comments | | Interactive /Public | | 50 | 39,749 |
| | | YouTube comments | | Interactive /Public | | 50 | 123,664 |
| | | Tweets | | Interactive/Public | | 50 | 452,165 |
| | | Reader commentary | | Interactive /Public | | 50 | 155,823 |
| | Mass Media (Burger 2000) | Reader's letter to the editor | | Public | Letter | 70 | 57,843 |
| | | News | | Public | News | 100 | 31,704 |
| | | Newspaper article | | Public | Article | 50 | 47,097 |
| | | Editorial | | Public | Editorial | 50 | 38,504 |
| | | Magazine article | | Public | Article | 50 | 57,843 |
| | | Commentary/opinion | | Public | | 50 | 46,346 |
| | Economics and Commerce (Hundt 2000) | Business communication | | Public | Invitations/ Letters/ | 50 | 132,182 |
| | Jurisprudence and the Legal System (Busse 2000) | Legal texts from the school and university | | Public | Schools laws/ Resolutions/ Regulations | 50 | 279,433 |
| | Everyday Use (Heinemann 2000b) | Recipe | | Public | Manual | 200 | 40,957 |
| | | Instruction manual | | Public | | 50 | 184,978 |
| | | Horoscope | | Public | | 50 | 40,626 |
| | | Job advertisements | | Public | | 105 | 30,271 |
| | Fiction Literature (Eroms 2008) | Prose | | Public | Novel/romance | 50 | 85,341 |
| | | | | Public | Youth literature | 50 | 274,865 |
| | Other | Non-fiction | | Public | | 50 | 207,243 |
| **Total Written** | | | | | | **1,875** | **9,088,632** |

Table 1 (continuation)

| Mode | Fields of Communication | Register | Sub-registers | Setting | Specific text-type included in corpus | Texts | Number of words |
|---|---|---|---|---|---|---|---|
| SPOKEN | Everyday Use (Heinemann 2000b) | Conversation | Pop/Rock/Pop-Rock/Hip Hop/Rap | Dialogue/Private | | 50 | 499,322 |
| | | Song lyrics | | Monologue/Scripted | | 150 | 49,369 |
| | Other | Institutional communication (FOLK) | | Dialogue/Public | | 50 | 319,111 |
| | | Interview (FOLK) | | Dialogue | | 23 | 205,662 |
| | | TED talk | | Monologue/Non scripted | | 30 | 53,055 |
| | Church and Religion (Simmler 2000) | Sermons | | Monologue/ Scripted | | 50 | 169,535 |
| | Mass Media (Burger 2000) | Spoken news | | Monologue/ Scripted | | 50 | 30,560 |
| | | Newspaper interview | | Monologue/ Scripted | Interview | 50 | 87,701 |
| | University and Scientific fields (Heinemann 2000b) | Academic speech | Oral exam | Monologue/Non-scripted | Exam | 38 | 167,735 |
| | | | Experts' Lecture (manuscript) | Monologue/Scripted | | 50 | 356,987 |
| | | | Experts' Lecture (transcription) | Monologue | Lecture | 14 | 87,032 |
| | | | Students' Lecture (transcription) | Monologue | | 29 | 134,862 |
| | Political Institutions (Klein 2000) | Political speech | | Monologue Scripted | | 50 | 90,271 |
| | Fiction Literature (Eroms 2008) | Theatre | | Dialogue scripted | | 50 | 279,286 |
| | | Films (Veirano Pinto 2013) | Documentary | Monologue/Dialogue scripted | | 50 | 253,428 |
| | | | Series | Dialogue scripted | Drama, Comedy, Romance | 38 | 110,808 |
| | | | Films | Dialogue scripted | Detective | 23 | 78,130 |
| | | | | Dialogue scripted | Comedy | 50 | 367,376 |
| | | | | Dialogue scripted | Drama | 50 | 281,587 |
| **Total Spoken** | | | | | | **897** | **3,621,817** |

Table 2: Spoken section of Koder (*Korpus deutscher Register*)

An important remark to be made is that this corpus is intended to be a monitor corpus. The design decisions described here refer to this first version which will be used to investigate register variation in contemporary German. More texts and registers will be added in the future according to necessity.

### *2.4. Processing*

After the material was collected, some processing was needed. All the text files were converted into text (.txt) format, either manually, by copying and pasting, or automatically, with the command pdftotext for PDF files. Sometimes, PDF files had to be manually corrected when they were converted into text format because the content became unreadable. Some transcripts from subtitles had an *l* instead of an *I* and vice versa in words as *Ich* which was then written *lch* or *als* which was written *aIs*. Contracted forms with *'s* were normalised to *es*. All these corrections were made using *sed*, a Unix utility which can be used for editing data (Kochan and Wood 2016: 70).

The files were cleaned semi-automatically: most texts had to be cleaned manually because the material to be removed was not uniform across texts. The texts from the spoken registers, chat and Wikipedia user talk had uniform material so that it was possible to use scripts written in Shell to clean them. The scripts are listed in Appendix 1.

## 3. CONCLUSION

This paper presented Koder, a multi-register-corpus of contemporary German which is composed of diversified register categories from a broad range of communicative situations. Thus, it comprises written and spoken texts as well as texts from computer mediated communication. The building of this corpus required decisions to be made not only about the number of texts and words but also about the number and type of registers to be included. The selection of the register categories was based on fields of communications (Brinker *et al.* 2000; Eroms 2008) and expanded with the addition of more registers from specific domains such as the Internet (Beißwenger and Lemnitzer 2013; Berber Sardinha 2014), films (Veirano Pinto 2013) and available corpora (DeReKo, *Folk and Gwiss, Dortmunder-Chat-Korpus* and *German Political Speeches*).

Decisions regarding the number of texts and words were guided by works of Biber (1990, 1993a, 1993b) and Berber Sardinha (2004).

The Koder corpus comprises a broad range of register categories which are used in the most diverse communicative situations by German speakers. Notwithstanding, some essential categories such as different types of television programmes could not be included in this first version of the corpus due to time and technical limitations. The expansion of the corpus in terms of registers, sub-registers and number of texts as well as a conducting further research will be considered in future research.

## REFERENCES


Adamzik, Kirsten. 2016. *Textlinguistik. Grundlagen, Kontroversen, Perspektiven*. Berlin: Mouton de Gruyter.

Barbaresi, Adrien. 2012. *German Political Speeches, Corpus and Visualization*. http://purl.org/corpus/german-speeches (15 November, 2016.)

Beißwenger, Michael. 2013. *Das Dortmunder Chat-Korpus: Ein Annotiertes Korpus zur Sprachverwendung und Sprachliche Variation in der Deutschsprachigen Chat-Kommunikation. LINSE*. http://www.linse.uni-due.de/tl_files/ PDFs/ Publikationen-Rezensionen/Chatkorpus_Beisswenger_2013.pdf (22 December, 2018.)

Beißwenger, Michael and Lothar Lemnitzer. 2013. Aufbau eines Referenzkorpus zur deutschsprachigen internetbasierten Kommunikation als Zusatzkomponente für die Korpora im Projekt "Digitales Wörterbuch der deutschen Sprache" (DWDS). *Journal for Language Technology and Computational Linguistics* 26/2: 1–22.

Berber Sardinha, Tony. 2004. *Lingüística de Corpus*. Barueri: Manole.

Berber Sardinha, Tony. 2014. 25 years later: Comparing Internet and pre-Internet registers. In Tony Berber Sardinha and Márcia Veirano Pinto eds., 81–105.

Berber Sardinha, Tony, Carlos Kaufmann and Cristina Acunzo. 2014. Dimensions of register variation in Brazilian Portuguese. In Tony Berber Sardinha and Márcia Veirano Pinto eds., 35–79.

Berber Sardinha, Tony and Márcia Veirano Pinto eds. 2014. *Multi-Dimensional Analysis, 25 years on: A Tribute to Douglas Biber*. Amsterdam: John Benjamins.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1990. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing* 5/4: 257–269.

Biber, Douglas. 1993a. Representativeness in corpus design. *Literary and Linguistic Computing* 8/4: 242–257.

Biber, Douglas. 1993b. Using register diversified corpora for general language studies. *Computational Linguistics* 19/2: 219–241.

Biber, Douglas and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Biber, Douglas, Mark Davies, James K. Jones and Nicole Tracy-Ventura. 2006. Spoken and written register variation in Spanish: A multi-dimensional analysis. *Corpora* 1/1: 1–37.

Brinker, Klaus, Gerd Antos, Wolfgang Heinemann and Sven Sager eds. 2000. Preface. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann and Sven Sager eds., XXIII–XXVIII.

Brinker, Klaus, Gerd Antos, Wolfgang Heinemann and Sven Sager eds. 2000. *Linguistics of Text and Conversation: An International Handbook of Contemporary Research*. *Volume 1*. Berlin: Mouton de Gruyter.

Burger, Harald. 2000. Textsorten in den Massenmedien. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann and Sven Sager eds., 614–628.

Busse, Dietrich. 2000. Textsorten des Bereichs Rechtwesen und Justiz. In Klaus Brinker Gerd Antos, Wolfgang Heinemann and Sven Sager eds., 658–675.

*Deutsches Referenzkorpus* (DeReKo), Wikipedia Diskussionen 2015. http://corpora.ids-mannheim.de/pub/wikipedia-deutsch/2015/ (20 November, 2017)

Eroms, Hans-Werner. 2008. *Stil und Stilistik: Eine Einführung*. Berlin: Schmidt.

Heinemann, Margot. 2000a. Textsorten des Bereichs Hochschule und Wissenschaft. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann and Sven Sager eds., 702–709.

Heinemann, Margot. 2000b. Textsorten des Alltags. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann and Sven Sager eds., 604–614.

Hundt, Markus. 2000. Textsorten des Bereichs Wirtschaft und Handel. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann and Sven Sager eds., 642–658.

IDS, *Datenbank für Gesprochenes Deutsch* (DGD), FOLK. http://dgd.ids-mannheim.de (9 October, 2019.)

IDS, *Datenbank für Gesprochenes Deutsch* (DGD), GWSS. http://dgd.ids-mannheim.de (9 October, 2019.)

Klein, Joseph. 2000. Textsorten in Bereich politischer Institutionen. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann and Sven Sager eds., 732–755.

Kochan, Stephen and Patrick Wood. 2016. *Shell Programming in Unix, Linux and OS X* (fourth edition). Indiana: Addison-Wesley.

Loewen, Shawn and Talip Gonulal. 2015. Exploratory factor analysis and principal components analysis. In Luke Plonsky ed. *Advancing Quantitative Methods in Second Language Research*. London: Routledge, 182–212.

Simmler, Franz. 2000. Textsorten des religiösen und kirchlichen Bereichs. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann and Sven Sager eds., 676–690.

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.

Veirano Pinto, Márcia. 2013. *A Linguagem dos Filmes Norte-americanos ao Longo dos Anos: Uma Abordagem Multidimensional*. São Paulo: PUC São Paulo dissertation.

Wiese, Ingrid. 2000. Textsorten des Bereichs Medizin und Gesundheit. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann and Sven Sager eds., 710–718.

Xiao, Richard. 2009. Multidimensional analysis and the study of world Englishes. *World Englishes* 28/4: 421–450.

*Corresponding author*
Andressa Costa
Praca Marechal Deodoro, 60
01150-010 São Paulo, Brazil
e-mail: acosta.andressa@gmail.com

received: August 2019
accepted: October 2019

# Appendix 1: Scripts used for the automatic cleaning of some texts

## EVERY DAY CONVERSATION and INSTITUTIONAL COMMUNICATION

```
cat filename.txt | grep [A-Za-z] | cut -f3 | sed -e 's/(.)//g' -e 's/([0-9]\.[1-9]*//g' -e 's/°hhh//g' -e 's/°hh//g' -e 's/flüstert//g' -e 's/°h//g' -e 's/°hh//g' -e 's/°hhh//g' -e 's/hhh°//g' -e 's/hh°//g' -e 's/h°//g' -e 's/räuspert sich//g' -e 's/räuspert//g' -e 's/lacht//g' -e 's/lachen//g' -e 's/+++//g' -e 's/schmatzt//g' -e 's/schnalzt//g' -e 's/seufzt//g' -e 's/hustet//g' -e 's/schluckt//g' -e 's/unverständlich//g' -e 's/schnieft//g' -e 's/Lachansatz//g' -e 's/Gemurmel,//g' -e 's/Gemurmel//g' -e 's/Blättern,//g' -e 's/Gelächter,//g' -e 's/Gelächter//g' -e 's/Geräusche,//g' -e 's/kichert//g' -e 's/stöhnt//g' -e 's/Nebengeräusche,//g' -e 's/Nebengeräusche//g' | tr -d '[]()' |

grep -v '^$'> filename_clean.txt
```

## INTERVIEW, ORAL EXAM and LECTURE

```
cat filename.txt | grep [A-Za-z] | cut -f3 | sed 's/(.)//g' | sed 's/([0-9]\.[1-9]*//g' | sed 's/°hhh//g'| sed 's/°hh//g' | sed 's/°h//g' | sed 's/hhh°//g' |sed 's/hh°//g' | sed 's/h°//g' | sed 's/räuspert sich//g' | sed 's/räuspert//g' | sed 's/lacht//g' | sed 's/+++//g' | sed 's/schmatzt//g' | sed 's/schnalzt//g' | sed 's/hustet//g' | sed 's/schluckt//g' | sed 's/unverständlich//g' | sed 's/lachend//g' | tr -d '[]()' | grep -v '^$'> filename_clean.txt
```

## CHAT

```
cat chattxt/ filename | grep -A1 '<messageBody>' chattxt/pc45.txt | tr '<' '\n' | grep -v '>' | cut -d'-' -f4 | grep -v '^$' >filename_clean.txt
```

## WIKIPEDIA USER TALK

```
cat wd_txt/filename | grep -A1 '<p>' wd_txt/file | tr '<' '\n' | tr -d '[]()' | grep -v '>' | cut -d'-' -f4 | grep -v '^$' >wd_clean/filename_clean.txt
```

RiCL Research in Corpus Linguistics

# The acquisition of L3 Spanish articles: What can be learned from a simple linear regression analysis?

Martin Testa
University of Warsaw / Poland

**Abstract** – Despite being extremely frequent in Spanish, articles pose one of the biggest challenges for Polish L1/English L2 students who learn it as a L3, even among the most advanced learners. However, little attention has been paid to the influence of the task at hand, especially since certain registers and writing styles tend to make use of a higher number of noun phrases, thus increasing both the amount of article syntactical environments and the probability of an error to be made. The main purpose of this article is to investigate whether there is a statistically significant relation between the number or article-related errors and an increasing number of article environments. In order to do this, a simple linear regression analysis was run for both B1 and B2 groups separately, and finally the whole level (n = 72). The results of the *t* tests suggest a significant linear relationship between the number of article-related errors and article tokens, between article-related errors and noun tokens, and between article and noun tokens. The article provides some considerations on possible actions to be taken both for researching and teaching the Spanish article system to [–ART] L1 learners.

**Keywords** – SLA; TLA; Spanish L3; article acquisition; simple linear regression

## 1. INTRODUCTION[1]

Although it may be true that articles might not be absolutely necessary in immediate environments (e.g. when ordering food; cf. Pica 1983), a more in-depth training of the semantic features of the article system seems necessary. This is especially the case when it comes to developing written competence, since in writing definiteness can only be expressed verbally.

[1] The author would like to express his very great appreciation to Prof. Jadwiga Linde-Usiekniewicz from the University of Warsaw for drawing attention to the issue of stylometric differences and their potential influence on the number of article-related errors. The author is also thankful to two anonymous reviewers for their thoughtful and constructive comments on an earlier version of this paper.

As will be shown in Section 2, below, evidence calls for a systematic pedagogical approach to the instruction of the article system, in which case learning the 'function' (cf. Pienemann 1998) of grammatical distinctions seems to be essential (cf. Widdowson 1988; Master 1994), especially at 'intermediate levels'. Some authors (cf. Master 1994; Ionin *et al.* 2004) argue that extensive exposure to comprehensible input – namely by means of the 'natural approach' (cf. Krashen 1982; Krashen and Terrell 1983) – seems not enough for L2 article acquisition. Master (1994) acknowledges, however, that this method is appropriate at the beginning levels of L2 instruction, at which point instruction of grammatical distinctions might be too detailed (cf. Allen 1983; Little 1994; Master 1997). More studies in terms of naturalistic acquisition are then required before definite conclusions can be drawn.

Some research has been carried out on the influence of pedagogical intervention in terms of English L2 article acquisition. For example, Master (1987: 116) suggests the following sequence for the assessment of article environments: [±count] > [±definite] > [±generic] > [±postmod] > [±proper] > [±idiomatic]. The facilitating effects of the application of this kind of systematic long-term instruction is supported by Ekiert and Han (2016), who have called for more conclusive research on the effectiveness of pedagogical intervention in L2 article instruction.

The situation is even less clear when it comes to other languages. For example, in the case of Spanish FL handbooks the article seems to be only treated at the initial stages and the understanding of its semantics is almost immediately taken for granted and subsequently abandoned (cf. Tarrés Chamorro 2002; Lin 2003; Hidalgo 2015).[2] This proves to be an inadequate approach given that even the formal and semantic differences between the article systems of two [+ART] languages are highly abstract and difficult to address, let alone for learners whose L1 is [–ART], i.e. Slavic languages.

In terms of linguistic typology, languages can be classified into 'article + tense' (i.e. Romance and Germanic languages) and 'aspect + case' languages (i.e. Slavic languages). Taking into account this classification, Polish and Spanish are typologically different (cf. Nowikow 2017). Polish differs from Spanish in that definiteness can be expressed by means of different strategies: namely, word order, aspect, grammatical case, lexical choice (i.e. the numeral 'one'), and deictic elements such as demonstratives or possessives (cf. Tarrés Chamorro 2002; Lin 2003; Shen Jie 2012; Hidalgo 2015;

[2] Some authors claim the same holds true for English (cf. Ionin *et al.* 2004; Sabir 2015).

Fernández Jódar 2017).[3] However, it should also be borne in mind that none of these strategies are completely equivalent to the semantics of articles (cf. Fernández Jódar 2017) and, as Ekiert and Han (2016: 151) note, even the very concept of *definiteness* "is often left to be inferred in a variety of ways." This is, for instance, the case in the Polish sentence *Potknąłem się o psa* ('I tripped over dog'). In this example, the lack of article does not give any hints as to whether the speaker is referring to *a dog* or *the dog* (cf. Allan 1986). In order to be more precise, for example, the speaker may either use the very name of the dog (i.e. *Potknąłem się o Reksia*), a determiner (i.e. *Potknąłem się o tego psa*), or fronting (i.e. *O psa się potknąłem*).

Studies such as Fernández Jódar (2006) have typically analyzed the ratio of article-related errors per one hundred words. However, it is still not clear whether a different approach to calculating the ratio of such errors would be more accurate, since different types of texts can be characterized by different writing styles (e.g. 'nominal' argumentative texts vs. 'verbal' narrative texts). The main purpose of this article is therefore to determine whether there exist statistically significant relations between: (a) the number of article-related errors and the number or article tokens; (b) the number of article-related errors and the number of noun tokens; and (c) the number of article tokens and the number of noun tokens.

The paper is organised as follows. Section 2 offers a review of the literature on article acquisition and instruction, which is intended to provide the reader with a basic understanding of the complexity of learning the article system, especially by learners whose mother tongue does not make use of articles: i.e. [–ART] L1 speakers. Section 3 provides a description of the data gathering procedures. The results will be presented in Section 4 and will be discussed in Section 5. Finally, Section 6 offers some concluding remarks and points out future lines of research.

## 2. LITERATURE REVIEW

There is extensive literature on the acquisition of English articles (cf. Master 1986, 1987, 1994, 2002; Ionin *et al.* 2004; Jaensch 2008; Yoo 2009; Sabir 2015; Ekiert and Han 2016; Sun 2016; or Şekerci Arıbaş and Cele 2019). As regards Spanish, some

[3] For a contrastive analysis between Spanish and Polish see Konieczna-Twardzikowa (1992), Tarrés Chamorro (2002) and Fernández Jódar (2017). For Spanish and Chinese see Lin (2003), Shen Jie (2012), and Hidalgo (2015).

research has been carried out on the acquisition of the Spanish article system by Polish L1 students (cf. Tarrés Chamorro 2002; Fernández Jódar 2006; Testa 2019, in press) and by Chinese L1 students (cf. Lin 2003; Shen Jie 2012; Hidalgo 2015).

Although it may be true that, as Master (1994: 230) note, "a person may communicate effectively in spoken English even when article use is entirely erroneous," article-related errors are especially visible in writing, where a weak command in the L2 article system might as well undermine the students' academic performance (cf. Master 1987, 1990, 1997). Many authors have pointed out that even high-intermediate level students use articles in an intuitive way (cf. Master 1994; Tarrés Chamorro 2002; Hidalgo 2015; Lema 2016). However, the way L1 speakers of [–ART] languages use articles is not random but actually responds to universal semantic distinctions (cf. Ionin *et al.* 2004, see section 2.1. below).

Despite being extremely frequent (cf. Hewson 1972), articles still pose great problems for students whose L1 does not have such a category (cf. Ringbom 1987; Master 1990, 2002; Mizuno 1999; Lin 2003; Park 2006; Harb 2014). These speakers often perceive articles as redundant (cf. George 1972) and frequently avoid their use (cf. Ringbom 2011), especially during the first stages of language learning (cf. Ringbom 2016). Jiang *et al.* (2011: 959, cited in Ekiert and Han 2016: 149) explain that grammatical morpheme acquisition is particularly demanding since its "related meaning is not grammaticalized in the learner's L1, which means that the related meaning is not part of the routinely activated meanings in the learner's mind." Master (1990) also points out that the article system is one of the latest aspects of L2 English to be fully acquired. In fact, as a result of lacking such a category in their native languages, [–ART] learners are one stage behind [+ART] learners when it comes to acquiring the L2 article system and therefore need more time (cf. Bailey *et al.* 1974; Hakuta 1976; Master 1997) to activate them in their mind.

Nonetheless, article-related errors are found even among the most advanced students (cf. Lin 2003; Ekiert and Han 2016) and also among students whose L1 features such a category (cf. Odlin 1989). Knowing an additional [+ART] L2 may indeed be of help, but only if the learners' command of the L2 is highly proficient or native-like, will their knowledge be available for transfer to the L3 (cf. Tarrés Chamorro 2002; Jaensch 2008; or Testa 2019, in press). At lower levels the students' mastery of the L2 article system may be far from perfect and, therefore, not very helpful. Moreover, not all

article systems are alike and both formal and conceptual differences between [+ART] languages add an extra layer of difficulty when it comes to acquiring the article system of an L3. Such a difficulty may even be further increased because of lexicalised and idiomatic uses in the article system of the different languages.

Harb (2014: 98–99) shows that students transfer the semantic properties of articles from the L1 to the L2. For example, this is the case between English [+ART] and Spanish [+ART] (cf. García Mayo 2008; Ionin *et al.* 2008; Isabelli-García and Slough 2012), and between English and Arabic [+ART] (cf. Kharma 1981; Schulz 2004). However, it remains unclear why [+ART] speakers are more able to transfer [+DEF]-usage than [–DEF]-usage into L2 English, given that languages such as Spanish and German also have indefinite articles (Master 1987). In this respect, Hidalgo (2015) draws attention to the fact that Mandarin Chinese may sometimes use the numeral *yī* ('one') in order to mark [–definite], but it may be omitted and for that reason positive transfer does not always take place.

Ionin *et al.* (2004) suggest two important hypotheses: 'the Article Choice Parameter' and the 'fluctuation hypothesis'. Regarding the Article Choice Parameter hypothesis, they suggest that [+ART] languages distinguish articles on the basis of either definiteness (i.e. English) or specificity (i.e. Samoan).[4] For example, on the one hand, Samoan features a marker for specificity *le* that corresponds to both English definite/specific *the* and indefinite/specific *a*. On the other hand, the Samoan marker for non-specific *se* may correspond to both English definite/non-specific *the* and indefinite/non-specific *a*. In fact, as Lyons (1999: 167) illustrates, definites can be either specific, as in *Joan wants to present the prize to the winner — but he doesn't want to receive it from her*, or non-specific, as in *Joan wants to present the prize to the winner — so she'll have to wait around till the race finishes*. Indefinites can also be either specific, as in *Peter intends to marry a/this merchant banker — even though he doesn't get on at all with her* or non-specific, as in *Peter intends to marry ~~a/this~~ merchant banker — though he hasn't met one yet*, as shown in Lyons (1999: 176). As Ionin *et al.* (2004: 17) point out, the Fluctuation Hypothesis holds that "L2 learners have full UG access to the two settings of the Article Choice Parameter" and that they fluctuate

[4] According to Ionin *et al.* (2004: 9), specificity involves "speaker intent to refer to an individual who exists in the actual world." This explains errors in [−definite, +specific] contexts, as illustrated in *I am visiting a/*the friend from college — his name is Sam Brown, and he lives in Cambridge now* (taken from Ionin *et al.* 2004: 41).

between those two settings "until the input leads them to set this parameter to the appropriate value."

Jenks (2018) provides a typology of definiteness marking. 'Bipartite' languages have one form of the definite article for unique contexts ($\text{DEF}_{\text{weak}}$) and a different one for anaphoric contexts ($\text{DEF}_{\text{strong}}$). An example is German, as provided in Schwarz (2009: 41): *In der Kabinettssitzung heute wird ein neuer Vorschlag vom Kanzler erwartet* ('In today's cabinet meeting, a new proposal by the chancellor is expected') vs. *In der Kabinettssitzung heute wird ein neuer Vorschlag von dem Minister erwartet* ('In today's cabinet meeting, a new proposal by the minister is expected'.) 'Marked anaphoric' languages, such as Mandarin, use bare nouns in unique contexts (Ø) but tend to use a demonstrative in anaphoric contexts ($\text{DEF}_{\text{strong}}$). Languages such as English and Cantonese are marked, i.e. both for unique and anaphoric contexts. Finally, Jenks explains that it seems theoretically impossible for a language to mark unique definites and not to have an anaphoric definite article, i.e. $\text{DEF}_{\text{weak}}$ for unique contexts and Ø for anaphoric ones.

We will now provide an overview in terms of LX English and LX Spanish article instruction research.

### *2.1. LX English article instruction*

Master (1987) studies the acquisition of the English article system by L1 speakers of Chinese (–ART), Japanese (–ART), Russian (–ART), Spanish (+ART) and German (+ART), in terms of four main categories of article use, namely generic (*a*, *the*, Ø), specific definite (*the*), specific indefinite (*a*, Ø) and ambiguous generic (*a*, Ø).[5] Master (1997) claims that both [–ART] and [+ART] groups show internal homogeneity, although it should be borne in mind that he had used only one subject per L1, per level, and at least some part of the results might be influenced by individual variation. For example, [–ART] groups tend to go from an overextended use of Ø at the lower levels (which is something expected, since their L1s have no articles), to an overextended use of *the* once they have become aware of the morpheme, and then to a slowly increasing use of the indefinite article *a*. On the other hand, [+ART] groups overuse *the* right from the start and use Ø correspondingly less than their [–ART] counterparts. The acquisition

[5] For examples in each category, see Master (1987: 23).

of the indefinite article seems to happen as an independent process (cf. Pica 1983, 1985; Master 1987). [+DEF]-flooding seems common at the intermediate levels, probably due to recognition of the fact that English noun phrases need a specifier, and once "they realize that Ø can be a specifier too, and they start to increase their Ø-usage" (cf. Master 1987: 88). [+DEF]-flooding at the intermediate levels has also been observed among Chinese L1 learners of Spanish (cf. Lin 2003; Shen Jie 2012; Hidalgo 2015). Jenks (2018: 501), for instance, points out that, in Mandarin, "unique definites are realized with a bare noun, [whereas] anaphoric definites are realized with a demonstrative, except in subject position" and that, while lacking a definite article, Mandarin seems to distinguish between unique and previously mentioned definites (i.e. anaphoric).

Master (1990: 461) suggests that the article system may be taught "as a binary division between classification (*a* and Ø) and identification (*the*)" and further argues that "[a]ll the other elements of article usage can be understood within this framework." Master (1990: 465) goes on to claim that "determining the correct article in English requires the simultaneous consideration of four features: 'definiteness' [±definite], specificity [±specific], countability [+count], and number [±singular]." According to Master (1990: 465), "number subset really only applies to [+count] nouns and should therefore only be considered a feature of that subset." This increases the students' cognitive load dramatically, and it is likely to slow down fluency considerably (cf. Krashen 1982). That is why Master (1990: 466) suggests "collapsing the features" [±definite] and [±specific] into one single feature [±identified], thus reducing the number of concepts to be taught to two: identification vs. classification.[6] Nevertheless, Master (1990: 474–476) acknowledges the complexity of article-usage when it comes to proper nouns and idiomatic phrases, which remain "in the realm of things which must be learned and memorized and for which there is rarely a productive rule." The binary framework, identification vs. classification, is tested in Master (1994: 245) who concludes that "the article system can indeed be learned and that it is perhaps the systematic presentation of the article system that makes the difference."

Master (1997) elaborates on the abovementioned binary framework and explores the differences between the two types of Ø (the zero article $Ø_1$, and the null article $Ø_2$). While $Ø_1$, (cf. *I ate <u>the pizza</u>* vs. *I ate <u>a pizza</u>* vs. *I ate <u>pizza</u>*), is "the most indefinite

[6] Master (1990: 469) notes that classification appeals to the awareness of the [±count] feature (*a* vs. Ø), as in *What's this? = It's a whiteboard/It's chalk/These are books/This is paper*.

article in the article system" (cf. Harb 2014: 91), $Ø_2$ (cf. *After dinner, I'm going home*) is the most definite one, because it is applied in contexts where "definiteness can still be internally attained without the addition of the article *the*" (cf. Harb 2014: 92). Moreover, Master (1997: 226) suggests that at beginning levels of L2 instruction only extensive exposure is advisable given that rules might be yet obscure to learners (cf. also Allen 1983; Little 1994). According to Master, at intermediate levels the article system should be taught in a systematic way by means of the binary framework (cf. Master 1990) or input processing (cf. VanPatten and Cadierno 1993). Finally, at more advanced levels a lexical rather than syntactic approach seems more appropriate, for instance, lexical minimal pairs Ø vs. *the* (cf. Master 1995).

Master (2002) finds that with the use of an information structure framework in one month of teaching the English article system to a group of students (mixed L1s, +Art and –Art), these made small but significant improvement when compared to a control group which had been taught by means of more traditional explanations, and a third group which had received no article instruction. Master (2002: 331) suggests that language teachers "present canonical information structure as a preliminary guess in determining the appropriate article for any noun, providing a further potential aid in learning the article system."

Ionin *et al.* (2004: 4) analyze article use by L1 Russian and L1 Korean (both [–ART] languages) learners of L2 English and argue that [+DEF]-flooding with indefinites "is systematic, being tied to the occurrence of the feature [+specific]," and that the same holds true for [–DEF]-flooding with definites (i.e. absence of the feature [+specific]). Ionin *et al.* (2004: 42) also point out that "as proficiency increases L2 learners are able to set the Article Choice Parameter (although many advanced learners still show fluctuation)." Their findings therefore suggest that there may be direct access to universal semantic distinctions in L2 acquisition, given that L1 transfer cannot account for access to the feature [+specific].

Jaensch (2008: 87) studies the acquisition of L3 German articles by native [–ART] Japanese speakers. She finds a positive effect of L2 English proficiency, since L1 Japanese/L2 English/L3 German learners seem to be "more aware of the definiteness feature – perhaps due to having acquired an L2 which has articles that mark definiteness in the same manner" – than [–ART] L1 Russian/L2 English and [–ART] L1 Korean/L2 English learners in Ionin *et al.* (2004).

Sabir (2015) analyzes article use by L1 Saudi (Hejazi) Arabic learners of L2 English in three tasks: article elicitation, acceptability judgment and elicited written production. Sabir finds no clear relation between explicit article instruction – i.e. instruction in definiteness, specificity and genericity, as well as translation activities – and article accuracy. Her results are consistent with the Fluctuation Hypothesis postulated by Ionin *et al.* (2004).

In line with Master (1987), Sun (2016) finds that, in her study, the [+ART] group produces much more correct article-related sentences than the [–ART] group. Sun (2016: 5) also finds that the Ø article "is the last one to be acquired and is the most difficult one for L2 learners of all levels," whereas the "indefinite article *a* is the first one and the easiest one for L2 learners of all levels to acquire." In fact, Sun claims that there is correspondence between the proficiency level and the acquisition of the indefinite article, although not the same can be claimed for the definite and Ø article.

Ekiert and Han (2016) analyze the acquisition of English articles by 65 Slavic learners (Polish n = 42, Russian n = 11, and Ukrainian n = 12) by using video retelling, missing word activities, and a translation methodology. Ekiert and Han (2016: 166) point out to that "[i]n general, the L1-Polish participants considered speaker-oriented identifiability sufficient for reference tracking in English" (i.e. givenness), thus having a harder time at choosing the right form of the article in the case of first-mention referents. In [+ART] languages, the definite article implies that the referent of the noun phrase must be identifiable either within the discourse or "uniquely identifiable to the hearer" (Birner and Ward 1994: 1), that is, there should be no room for ambiguity in the ears of the hearer (cf. Hawkins 1991). Following Trenkic (2002), Ekiert and Han (2016: 165–166) explain that in Slavic languages, "the way objects are referred to mirrors the state of knowledge of the speaker only," and this often results in interference at the level of discourse. Following Yoo (2009), Ekiert and Han (2016: 150) point out that, although most student textbooks put emphasis on the anaphoric discourse-oriented use of *the* (i.e. *I bought a book. On Tuesday*, *I finished <u>the</u> book*), it is the cataphoric and situational uses of the definite article which are much more common in conversational English (i.e. *I had an argument at <u>the</u> office*).

Finally, Şekerci Arıbaş and Cele (2019) compare the performance of L1 Turkish/L2 German learners of L3 English and L1 Turkish leaners of L2 English, and

find that L3 English learners are significantly more accurate than L2 English learners in all article contexts. Their results are also consistent with the Fluctuation Hypothesis.

In this section, we have seen that many studies support the claim that knowledge of article semantics can be transferred from the L1 to the target language, as well as from a L2 [+ART] to an L3, especially if the level in the L2 is high. Proficiency in the target language has also been linked to a better command of article use. Moreover, [+DEF]-flooding with indefinites and [–DEF]-flooding with definites have been found to be systematic and their motivation can be linked to a readjustment of the Article Choice Parameter. Therefore, although there is some evidence on the benefits of systematic explicit instruction, there is also evidence in support of the Fluctuation Hypothesis.

### *2.2. LX Spanish article instruction*

Because of the perceived differences between [–ART] and [+ART] languages, Spanish articles pose a problem for Polish L1 students, regardless of the task and the level (cf. Fernández Jódar 2006: 97). Fernández Jódar (2006) identified up to twenty types of article-related errors, the most frequent being the unnecessary use of the definite article (e.g. *come * la carne*).[7] Furthermore, Fernández Jódar (2006: 106) concludes that articles cannot be said to be fully acquired by Polish L1 students (cf. Tarrés Chamorro 2002), and that the greatest problem lies in the unnecessary use of Spanish articles with non-referential noun phrases and the incorrect avoidance of the article, particularly the definite one. On the other hand, Fernández Jódar (2006: 107) notes progress in the use of the zero article with proper nouns and idioms, as well in the use of the neutral article *lo*, which is absent at early levels of instruction. However, it is not clear whether this represents a better command of the semantics of the zero article or it is rather correct avoidance of the article, that is, a covert error.

Tarrés Chamorro (2002: 79) also argues that a high level in a [+ART] L2 might be beneficial, although the evidence is sometimes contradictory. Moreover, he argues that L2 English does not seem to be of much help, probably because students have not yet fully acquired the English article system either. Tarrés Chamorro (2002: 74) reports the

[7] Pragmatics and semantics are crucial to determining the correctness of a sentence like *come (*) la carne*. The definite article is used in imperative sentences as in *¡Come la carne!* (‘eat your meat!’). However, in declarative clauses, the definite article is not used: that is, when referring to the act of eating meat regularly *Mi hijo no come carne*. (‘My son doesn’t eat meat.’)

case of an L1 Polish student, who had been in permanent and intense contact with the target language, and seemed to have achieved native-like control of the L3 Spanish definite article, mostly because of her extraordinarily high motivation and engagement. This could show that higher levels of motivation and engagement might be linked to more exposure to the target language, as well as to more attention to its distinctive features.

Lin (2003) studies 80 writing assignments by Chinese L1 (Taiwanese) multilingual learners of Spanish from four different levels at the University of Alcalá (Spain), and analyzes article-related errors in terms of *signifier* and *signified*-related errors (i.e. article omission; unnecessary use of article; and wrong choice of article). As a general strategy, Lin detects initial article omission to subsequent [+DEF]-flooding, and also finds that the most common type of error is article omission (especially [+DEF]-omission). At lower levels, Lin (2003) demonstrates that INDEF usage seems to be aided by the transfer of the Mandarin Chinese numeral *yī*).[8] Lin argues that the use of the Chinese numeral should result in positive transfer and, for this reason, the number of Ø-for-INDEF errors should be low. However, the Chinese numeral can be often omitted and, as a consequence, interference is still possible and positive transfer does not often take place (see also Hidalgo 2015). On the other hand, it has been shown (cf. Lapesa 2000; Lin 2003; Hidalgo 2015) that Mandarin allows the numeral to be placed before 'classifying' predicate nouns while Spanish does not (especially after copular verbs), thus resulting in interference, as in **soy un estudiante* ('I am a student') or *esto es *un asunto mío* ('this is my issue'). Once again, pragmatics and semantics are crucial in determining the correctness of these utterances. For example, the zero article is used after copular *ser* ('to be') as is the case with professions, cf. *Marta es actriz* ('Marta is an actress'), but the definite article can be used in emphatic contexts *Marta es la actriz más famosa* ('Marta is the most famous actress'), and the indefinite article may be used for comparison *Marta es una actriz* ('Marta {behaves ~ looks ~ talks} like an actress').

Shen Jie (2012) analyzes 90 writing assignments from A1-B2 Chinese L1 students of Spanish at the 'Official School of Languages' in Barcelona, and finds that the major issue for the students seemed to be the [±definite] feature, as well as the use of the Ø article when it comes to bare nouns.

[8] This is also noted in the case of Polish L1 learners (cf. Fernández Jódar 2006). We argue that this is due to the fact that students seem unaware of the rule that, with the exception of topicalized unaccountable nouns, Spanish nouns in subject position require the use of the article.

Hidalgo (2015) studies 20 article-related tasks (11 fill-in-the-blank and 9 translation exercises) from Chinese L1 learners of Spanish at B1 level, and finds that the most frequent errors to be DEF-for-INDEF ([+DEF]-flooding) as well as INDEF-for-DEF, followed by Ø-for-DEF.[9]

Finally, Testa (in press) explores whether knowledge of a [+ART] L2 was of any help when learning the article system of L3 Spanish. In order to do this, he analyzes both the production and comprehension of 25 Polish L1 intermediate-level students (B1 n = 11, B2 n = 14) by means of a translation task, a grammaticality judgment task, and post-task interviews. When it comes to the translation task, Spanish generic definite (Polish Ø) and bare nouns after copular *ser* ('to be') turned out to be the most difficult sentences to translate, with less than 40% of target-like answers each in both groups. In the case of the grammaticality judgment task, both groups find it particularly hard to recognize wrong Ø use (cf. George 1972) and, in contrast, have very little difficulty to correct wrong DEF-for-Ø and INDEF-for-Ø. Finally, the post-task interviews reveal that in many cases students tend to choose between article forms intuitively. They may be aware of some of the semantic/pragmatic distinctions but often fail at mapping them onto the right article form. Furthermore, during the post-task interviews none of the students pointed to English L2 influence, except for one student with native-like command of English. On the other hand, other languages (i.e. German, Italian and Portuguese) are considered to have an influence on the students' choices, which frequently results in interference (cf. Testa 2019: 40–41). The results echo Tarrés Chamorro's (2002: 51) suggestion that positive transfer might happen only if the level in the L2 is very high, that is, proficient user or native-like.

In this section we have observed that [+DEF]-flooding is also common in L3 Spanish as a general strategy by [–ART] learners, following initial article omission. Moreover, high proficiency in an [+ART] L2 seems to help in transferring article-related knowledge into L3 Spanish.

To the best of our knowledge, there are currently no studies that examine whether a higher number of article-related errors can be expected with increasing article usage. In other words, the type of task may exert an influence on the total number of article-related errors since certain writing styles tend to make a more frequent use of noun phrases – i.e. 'nominal' argumentative texts, as opposed to 'verbal' narrative texts) –

[9] This is especially the case with the verb *tener* ('to have'), as shown in Testa (2019: 273, 284, 295).

thus increasing both the amount of article environments and the probability of error. When it comes to Spanish, Fernández Jódar (2006) has analyzed the number of article-related errors per one hundred words but, as stated above, we may expect to find more articles and nouns in argumentative texts and, therefore, more environments for article-related errors. We argue that such an analysis would be a better indicator of the real ratio of article-related errors than the number of errors per one hundred words, or else the mean average of such errors.

In light of these considerations, the present study will address the following research questions:

1. Is there a significant relationship between the number of article-related errors (ArE) and the number of article tokens (AT)?
2. Is there a significant relationship between the number of article-related errors (ArE) and the number of noun tokens (NT)?
3. Is there a significant relationship between the number of article tokens (AT) and the number of noun tokens (NT)?

## 3. THE STUDY

The current study uses sample data collected from two intermediate-level (B1 n = 35; B2 n = 37) Spanish language courses that took place at the Faculty of Modern Languages at the University of Warsaw, between December 2016 and June 2017. The sampled population was highly homogeneous, since the participants were all multilingual Polish L1 university students doing a degree in Linguistics, with ages ranging from 19 to 25. They had English as their second language and were studying Spanish either as L3 (in most cases) or L4. The students had been admitted for the B1 and B2 courses on the basis of having passed an A2 and a B1 course, respectively.

The corpus consists of 72 compositions in total (15,288 words) coming from three different writing assignments about topics that had been discussed orally in class, in order to reduce the possible amount of errors due to ignorance of vocabulary. The topics are: 1) *Mis mejores vacaciones* (‘The best holidays I have ever had’) (B1 n = 13; B2 n = 15); 2) an e-mail to a friend telling them about a personal anecdote (B1 n = 10; B2 n = 10); and 3) *Ventajas y desventajas del uso de Internet a nivel educativo* (‘Advantages

and disadvantages of using Internet for educational purposes') (B1 n = 12; B2 n = 12). All assignments were written in class and students were given a total of forty-five minutes to do them. The first assignment took place at the end of the winter semester (January 2017), the second one was carried out before the Easter break (April 2017), and the third assignment was scheduled at the end of the summer semester (June 2017). At no time were the students told that the activities were aimed at analyzing their use of Spanish articles. Moreover, since the data come from two narrative texts and one argumentative text there might be stylometric differences with regard to the number of article environments. For example, in the narrative texts a higher number of verbs should be expected, whereas in the argument text more noun phrases could be used.

In order to answer our three research questions a simple linear regression analysis was carried out. A simple linear regression analysis is normally used in order to determine whether the resulting regression line fits the data better than the average mean line. In our study, we will try to determine whether there is a significant relationship between: (a) the number or article-related errors (ArE) and article tokens (AT); (b) the number of article-related errors (ArE) and noun tokens (NT); and (c) between the number of article tokens (AT) and noun tokens (NT). If the results are significant, we may conclude that those relations can predict the amount of article-related errors better than the mean average number of errors (our first two research questions), as well as predicting the amount of articles better the mean average number of article tokens (our third research question). The equations used in the analysis are shown in Table 1.

| Regression line | Slope calculation | Intercept calculation | Coefficient of determination |
|---|---|---|---|
| $y_i = b_0 + b_1 x_i$ | $b_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$ | $b_0 = \overline{y} - b_1 \overline{x}$ | $r^2 = \frac{SSR}{SST}$ |
| **SST** | **SSE** | **SSR** | **t test** |
| $\sum(y_i - \overline{y})^2$ | $\sum(y_i - y_i)^2$ | $SSR = SST - SSE$ | $t = \frac{b_1}{s_{b_1}}$ |
| **Standard error** | **Confidence intervals** | **Standar deviation of y *** | **Prediction intervals** |
| $\sigma = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}$ | $y* \pm t_{a/2} s_{y*}$ | $s_{y*} = s\sqrt{\frac{1}{n} + \frac{(x* - \overline{x})^2}{\sum(x_i - \overline{x})^2}}$ | $y* \pm t_{\alpha/2} s_{pred}$ |

Table 1: Equations used in the simple linear regression analysis

## 4. RESULTS

### *4.1. Regression analysis of groups B1 and B2 (separately)*

#### 4.1.1. Number of article-related errors (ArE) vs. Number of article tokens (AT)

As can be noticed in Table 2 below, the regression line for group B1 ($y_i$ = 0.2242$x$ + 3.7771) suggests that for every (1) article environment ($x$) we would expect the amount of article-related errors to increase by 0.2242. In Figure 1, it can be seen how the mean line relates to the observed number of errors on the scatter plot. The coefficient of determination ($r^2$) reveals that almost one third of the total sum of squares can be explained by using the estimated regression equation to predict the amount of article-related errors. In other words, as much as 30.78% of the variance in the amount of article-related errors can be explained by increasing article use. The standard error value ($\sigma$ = 3.5611) was obtained from the square root of the value resulting from the division of the SSE (418.4808) by the degrees of freedom ($n$ – 2 = 33). The confidence interval for the slope was estimated at (0.1055, 0.3429) and it does not contain zero, so the null hypothesis – that is, the mean average number of article-related errors fits the data better – can be rejected.[10] Finally, the t-test indicates that there is a statistically significant relationship between the number of article-related errors and the number of article tokens (t = 3.8324 > $t_{crit}$ = 2.03).

[10] All calculations in this study use $\alpha$ 0.05.

| | ArE vs. AT | | ArE vs. NT | | AT vs. NT | |
|---|---|---|---|---|---|---|
| | B1 (df = 33) | B2 (df = 35) | B1 (df = 33) | B2 (df = 35) | B1 (df = 33) | B2 (df = 35) |
| $y_i$ | 0.2242$x$ + 3.7771 | 0.193$x$ + 5.1755 | 0.1612$x$ +0.4559 | 0.1971$x$ – 0.5097 | 0.4906$x$ – 3.5209 | 0.2967$x$ – 7.8495 |
| $\overline{x}$ | 20.7429 | 24.2973 | 49.4571 | 55.4595 | 49.4571 | 55.4595 |
| $\overline{y}$ | 8.4286 | 9.8649 | 8.4286 | 9.8649 | 20.7429 | 24.2973 |
| $\sum (x_i - x)^2$ | 37000.6871 | 4929.7297 | 8537.6871 | 19679.1892 | 8536.6871 | 19679.1892 |
| *SSE* | 418.4808 | 982.678 | 382.7325 | 477.6808 | 1645.9889 | 3197.2757 |
| *SSR* | 186.0906 | 183.6463 | 221.8389 | 688.6435 | 2054.6968 | 1732.4540 |
| *SST* | 604.5714 | 1166.3243 | 604.5714 | 1166.3243 | 3700.6857 | 4929.7297 |
| $r^2$ | 0.3078 | 0.1575 | 0.3669 | 0.5904 | 0.5552 | 0.3154 |

Table 2: Regression lines for groups B1 and B2

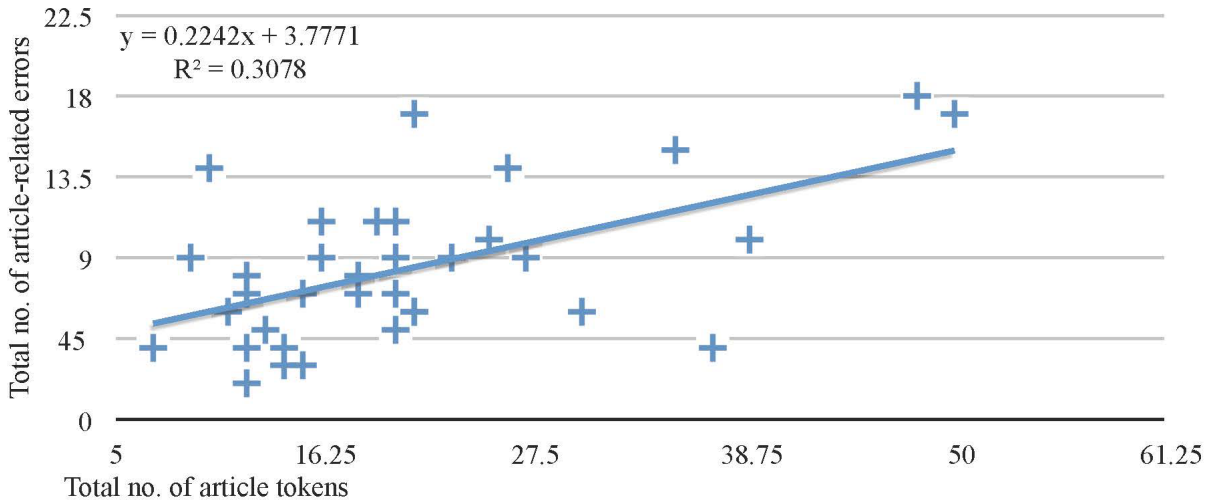


Figure 1. Article-related errors (ArE) vs. Article tokens (AT) (B1)

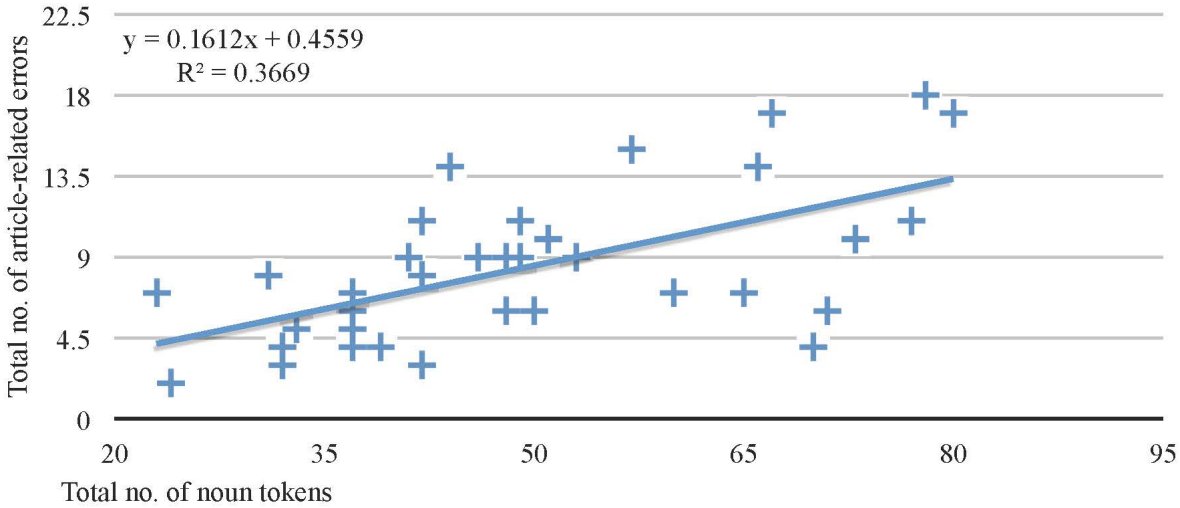


Figure 2: Article-related errors (ArE) vs. Noun tokens (NT) (B1)

With regard to group B2, the calculated regression line ($y_i = 0.193x + 5.1755$) reveals a lower frequency of article-related errors per article token, as shown in Figure 2, above. The coefficient of determination ($r^2$) suggests that the regression line is not a good fit, since only 15.75% of the variance in the amount of article-related errors can be explained by the use of articles itself. The standard error value ($\sigma = 5.2987$) was obtained from the square root of the value resulting from the division of the SSE (982.678) by the degrees of freedom ($n - 2 = 35$). There is confidence that the actual slope lies within the interval (0.0397, 0.3463) and it does not contain zero, so the null hypothesis can also be rejected. Finally, the *t*-test also reveals a significant relationship between the variables ($t = 2.5563 > t_{crit} = 2.03$).

### 4.1.2. Number of article-related errors (ArE) vs. Number of noun tokens (NT)

Because we are dealing with the same amount of article-related errors, the mean values for *y* and *SST* remain the same (cf. Table 2, above).

The regression line for group B1 was calculated as ($y_i = 0.1612x + 0.4559$), which suggests that for every new noun ($x$) we would expect the amount of article-related errors to increase by 0.1612 (cf. Figure 3). The coefficient of determination ($r^2$) reveals that up to 36.69% of the variance in the amount of article-related errors can be explained by increasing noun use. The confidence interval for the slope was estimated at (0.0864, 0.236) and it does not contain zero, so the null hypothesis can again be rejected. Furthermore, the *t*-test suggests a significant linear relationship between the number of article-related errors and the number of noun tokens ($t = 4.3686 > t_{crit} = 2.03$).

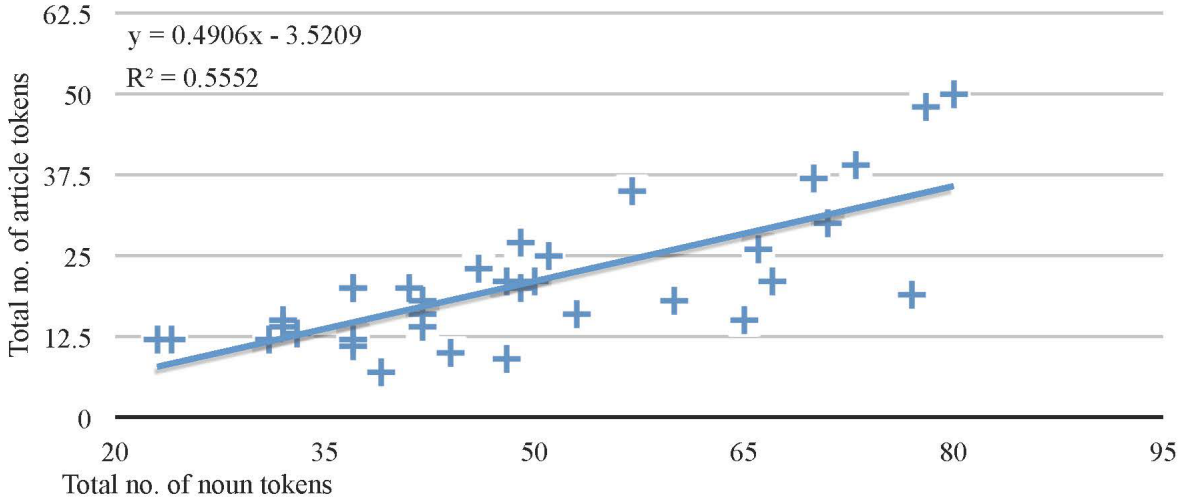


Figure 3: Article-related errors (ArE) vs. Noun tokens (NT) (B1)

With respect to group B2, the regression line ($y_i = 0.1871x + 0.5097$) suggests a slightly higher frequency of article-related errors per noun token, as shown in Figure 4. The coefficient of determination as ($r^2$) suggests a fairly good fit given that up to 59.04% of the variance in article-related errors can be explained by the number of nouns in the writing assignments. There is confidence that the slope lies within the interval (0.1336, 0.2407), and the null hypothesis can also be rejected as the interval does not contain zero. The *t*-test confirms a statistically significant linear relationship between the variables ($t = 7.1141 > t_{crit} = 2.03$).

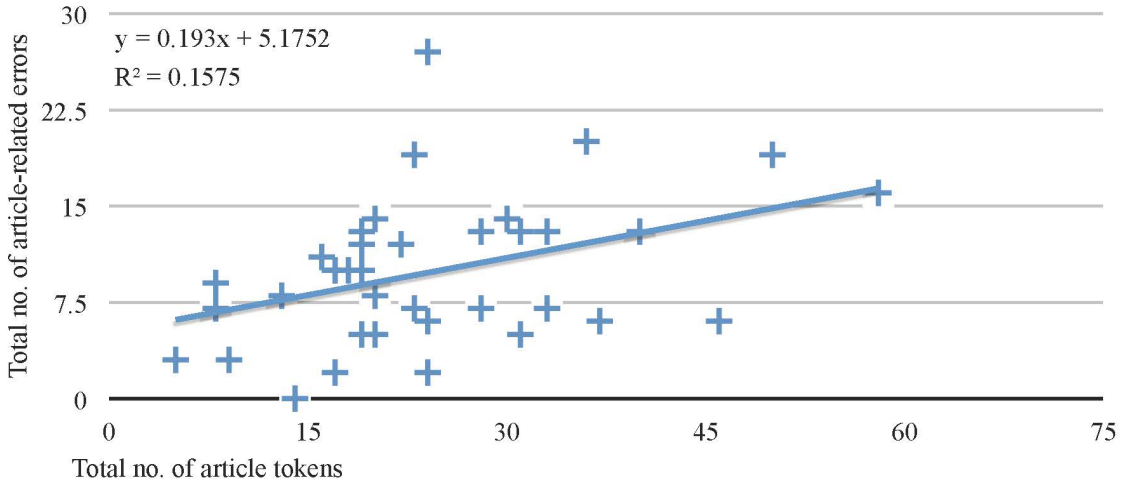


Figure 4: Article-related errors (ArE) vs. Article tokens (AT) (B2)

4.1.3. Number of article tokens (AT) vs. Number of noun tokens (NT)

The regression line for group B1 ($y_i = 0.4906x + 3.5309$) suggests for every new noun ($x$) the amount of articles is expected to increase by 0.4906, as shown in Figure 5, below. The coefficient of determination ($r^2$) indicates that up to 55.52% of the variance in the use of articles can be explained by the number of nouns used in each writing assignment. There is confidence in the interval (0.3354, 0.6458) containing the true slope of the regression line, and because the interval does not contain zero, the null hypothesis can be rejected. The *t*-test reveals a statistically significant relationship between the number of article tokens and noun tokens ($t = 6.4215 > t_{crit} = 2.03$).

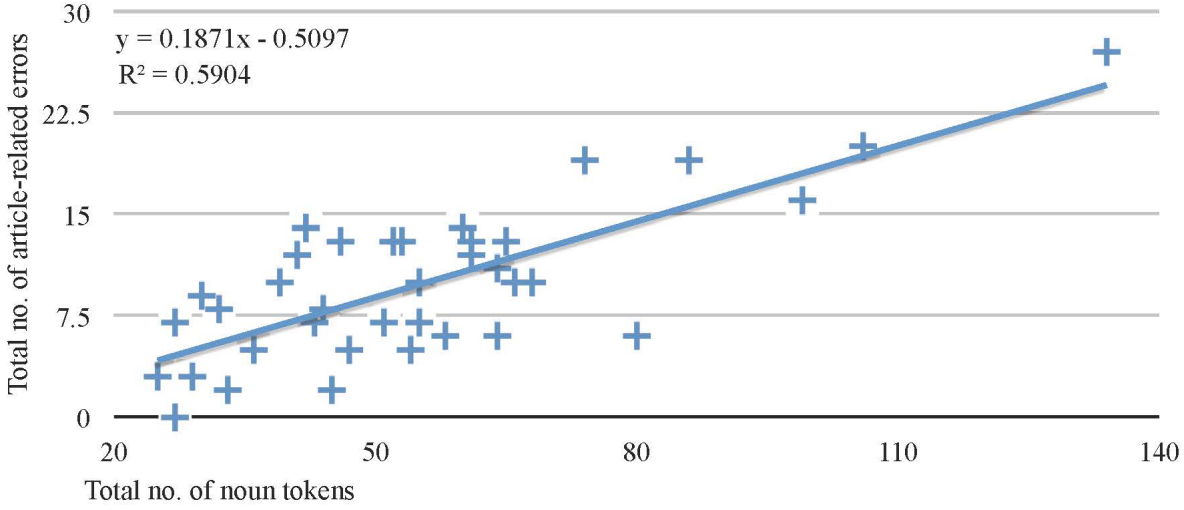


Figure 5: Article-related errors (ArE) vs. Noun tokens (NT) (B2)

As for group B2, the estimated regression line is ($y_i$ = 0.2967$x$ + 7.8425) and this suggests a lower frequency of articles per noun token, as illustrated in Figure 6. However, the coefficient of determination ($r^2$) does not suggest a good fit, since only 35.14% of the variance in article tokens can be explained by the amount of nouns. There is confidence that the slope lies within the interval (0.1584, 0.435), and since the interval does not contain zero, the null hypothesis can again be rejected. The result of the *t*-test confirms that the relationship between the variables is statistically significant ($t = 4.3568 > t_{crit} = 2.03$).

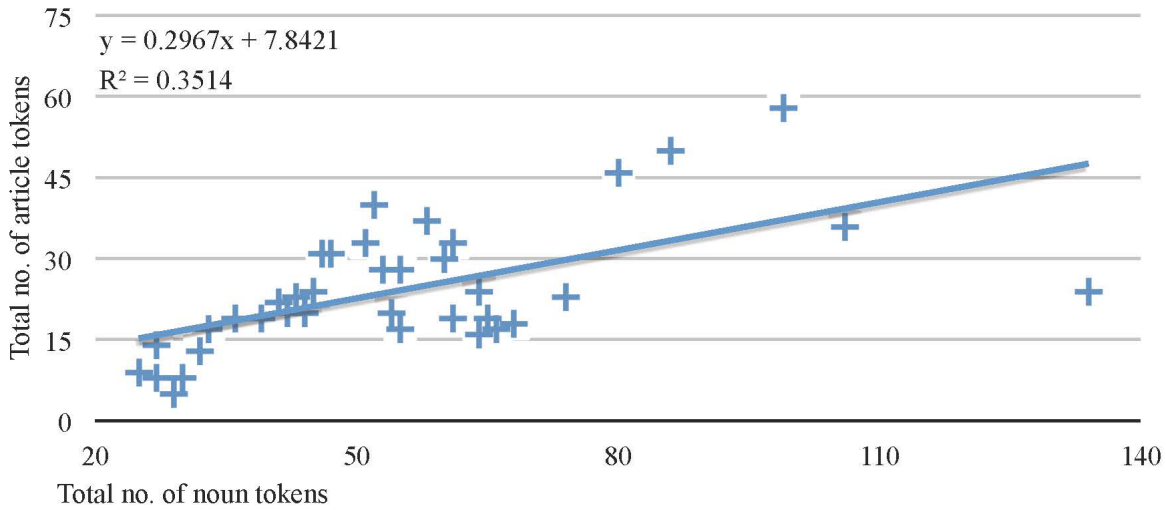


Figure 6: Article tokens (AT) vs. Noun tokens (NT) (B2)

4.1.4. Confidence intervals for a specific value of article and noun tokens ($x*$)

Finally, confidence intervals have been generated for the two groups separately in order to estimate: (a) a mean value of article-related errors every $x^*$ article tokens; (b) a mean value of article-related errors every $x^*$ noun tokens; and (c) a mean number of articles every $x^*$ noun tokens. However, it should be stressed that confidence intervals were calculated for the rounded up value of $\overline{x}$ in order to get predicted average mean value of $y^*$ for a real value of $x^*$.[11]

A) Group B1

Firstly, by applying the regression equation is ($y_i$ = 0.2242$x$ + 3.7771) an average mean of 8.4853 article-related errors is estimated for every 21 article tokens. The confidence intervals was then calculated by adding the margin of error ($\pm t_{a/2} s_{y*}$), and it predicts that 95% of the sampled population in group B1 would produce a mean between 7.2624 and 9.7082 article-related errors for every 21 article tokens.

Secondly, the regression model ($y_i$ = 0.1612$x$ +0.4559) suggests an average mean of 8.3547 article-related errors for every 50 noun tokens has been estimated. According to the confidence interval, 95% of the sampled population are predicted to make a mean between 7.1849 and 9.5245 article-related errors for every 50 noun tokens.

Thirdly, according to the regression line ($y_i$ = 0.4906$x$ – 3.5209), an average mean of 21.0091 articles is estimated for every 50 noun tokens. The generated confidence interval estimate that 95% of the sampled population will produce a mean between 18.5831 and 23.4351 articles for every 50 noun tokens. However, it should be borne in mind that this is an estimate of article use and does not represent correct article use.

[11] Although it would have been more precise to calculate it for the $x$ value, it would not have yielded an accurate picture. This occurs because no student produces one noun/article and a half. Therefore, values for $x$* have been rounded up.

B) Group B2

Firstly, the regression equation ($y_i$ = 0.193$x$ + 5.1755) predicts an average mean of 10.0005 article-related errors for every 25 article tokens. The confidence interval was then calculated by adding the error margin, and it predicts that 95% of the sampled population in group B2 would produce a mean between 8.8298 and 11.1712 article-related errors for every 25 article tokens.

Secondly, the regression model ($y_i$ = 0.1871$x$ – 0.5097) predicts an average mean of 9.9679 article-related errors every 56 noun tokens. According to the confidence interval, 95% of the participants are expected to produce a mean between 8.7353 and 11.2005 article-related errors for every 56 noun tokens.

Finally, the regression line ($y_i$ = 0.2967$x$ + 7.8425) predicts an average mean of 24.4577 articles for every 56 nouns. The generated confidence interval estimate that 95% of the sampled population will produce a mean between 21.2687 and 27.6467 articles for every 56 nouns in the B2 group. Once again, this represents an estimate of article use but does not represent correct article use.

*4.2. Regression analysis of the intermediate Level Group (B1+B2)*

4.2.1. Number of article-related errors (ArE) vs. Number of article tokens (AT)

The calculated regression line ($y_i$ = 0.2115$x$ + 4.3938) suggests that for every new article environment ($x$), we would expect the amount of error to increase by 0.2115, as illustrated in Figure 7, below. The coefficient of determination ($r^2$) reveals that only 21.91% of the variance in article-related errors can be explained by using the number of article tokens. The confidence interval (0.1318, 0.2912) does not contain zero so the null hypothesis (the mean average number of article-related errors fits the data better) can be rejected. Furthermore, the *t*-test reveals a significant linear relationship between the number of article-related errors and the number of article tokens ($t = 4.4340 > t_{crit} = 1.67$).

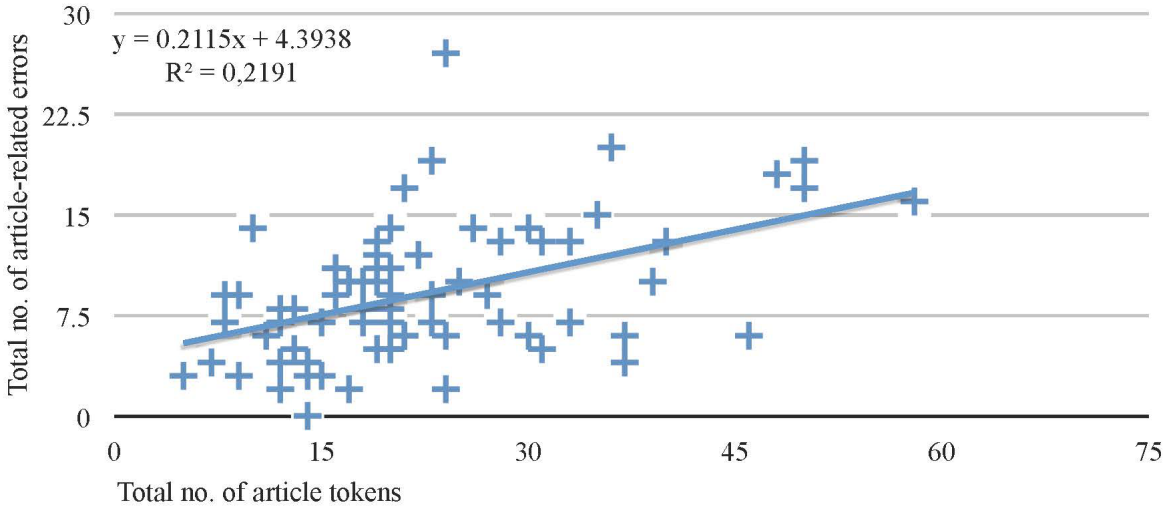


Figure 7: Article-related errors (ArE) vs. Article tokens (AT) (B1+B2)

4.2.2. Number of article-related errors (ArE) vs. Number of noun tokens (NT)

Due to the fact that we are dealing with the same amount of ArE, the mean values for $y$ remain the same, as illustrated in Table 3, below. In this case, the regression line ($y_i$ = 0.1806$x$ – 0.3218) suggests that for every new noun ($x$) the amount of article-related errors is expected to increase by 0.1806, as can be seen in Figure 8. The coefficient of determination ($r^2$) indicates a fairly good fit, since up to 52.06% of the variance in article-related errors can be explained by the number of nouns in each composition. There is confidence that the slope lies within the interval (0.146, 0.2152), and since it does not contain zero, the null hypothesis can then be rejected. In addition, the $t$-test confirms a statistically significant linear relationship between the variables ($t = 8.7346 > t_{crit} = 1.67$).

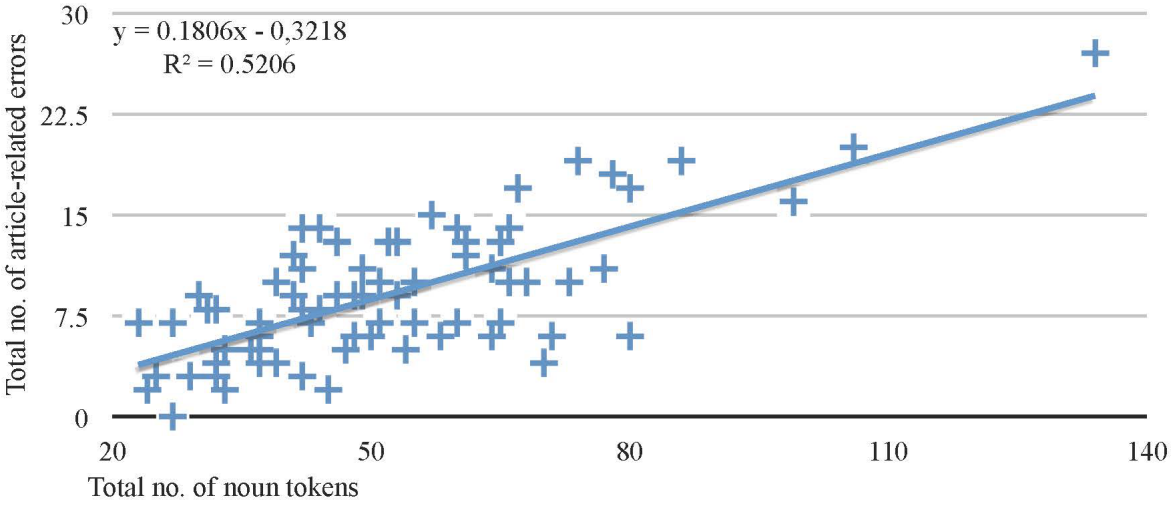


Figure 8: Article-related errors (ArE) vs. Noun tokens (NT) (B1+B2)

| | ArE vs. AT | ArE vs. NT | AT vs. NT |
|---|---|---|---|
| $y_i$ | 0.2115$x$ + 4.3938 | 0.1806$x$ – 0.3218 | 0.3607x + 3.6184 |
| $\overline{x}$ | 22.5694 | 52.5417 | 52.5417 |
| $\overline{y}$ | 9.1667 | 9.1667 | 22.5694 |
| $\sum (x_i - x)^2$ | 8857.6528 | 28863.8750 | 28863.8750 |
| *SSE* | 1411.8733 | 866.6796 | 51202.6274 |
| *SSR* | 396.1267 | 941.3204 | 37755.0253 |
| *SST* | 1808 | 1808 | 8857.6528 |
| $r^2$ | 0.2191 | 0.5206 | 0.4239 |

Table 3: Regression lines for the Intermediate Level group (n = 72)

4.2.3. Number of article tokens (AT) vs. Number of noun tokens (NT)

In this case, the resulting regression line ($y_i$ = 0.3607$x$ + 3.6184) indicates that for every new noun ($x$) the amount of article tokens is expected to increase by 0.3607 (cf. Figure 9). The coefficient of determination ($r^2$) suggests that 42.39% of the variation in the number of article tokens can be explained by the number of nouns in each composition. There is confidence in the interval (0.2768, 0.4446) containing the true slope of the regression line. The null hypothesis can be rejected because the interval does not contain zero. Finally, the *t*-test reveals a significant relationship between the variables (t = 7.1853 > $t_{crit}$ = 1.67).

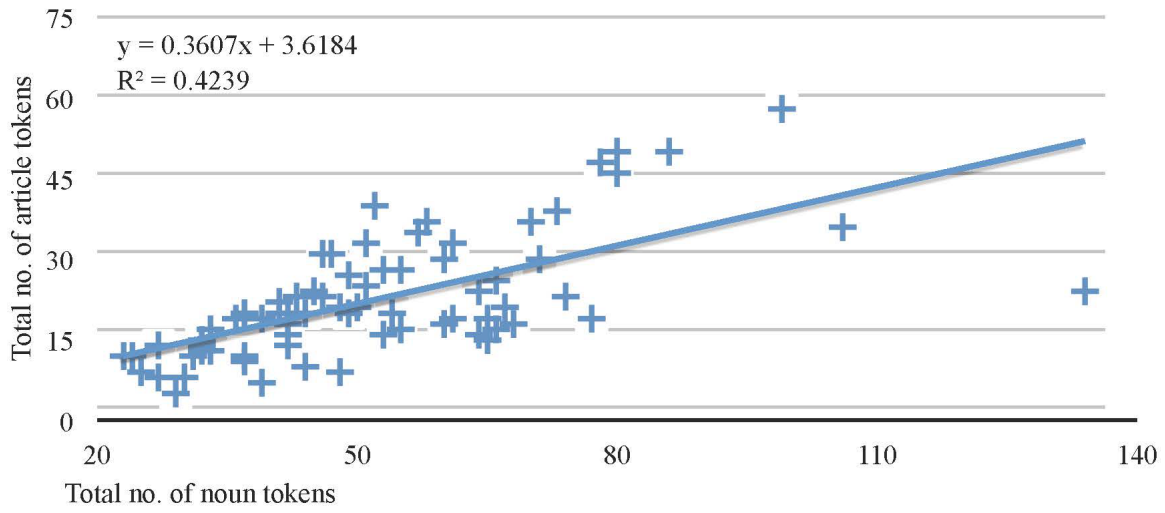


Figure 9: Article tokens (AT) vs. Noun tokens (NT) (B1+B2)

4.2.4. Confidence intervals for a specific value of article and noun tokens ($x$*)

Finally, confidence intervals have been generated in order to estimate: (a) a mean value of article-related errors every $x$* article tokens; (b) a mean value of article-related errors every $x$* noun tokens; and (c) a mean number of articles every $x$* nouns. However, it should be stressed that confidence intervals were calculated for the rounded value (cf. note 11, above) of $\overline{x}$ in order to get predicted average mean value of y* for a real value of $x$*.

Firstly, according to the regression model ($y_i$ = 0.2115$x$+ 4.3938), an average mean of 9.2583 article-related errors is predicted for every 23 article tokens. After adding the margin of error ($\pm t_{a/2} s_{y*}$), the confidence interval indicates that 95% of the sampled population in the Intermediate Level group was predicted to produce a mean between 8.3734 and 10.1432 article-related errors for every 23 article tokens.

Secondly, the regression model ($y_i$ = 0.1806$x$ – 0.3218) predicts that the students will produce a mean value of 9.25 article-related errors for every 53 noun tokens. Moreover, the confidence interval suggests that 95% of the students will produce a mean between 8.557 and 9.943 article-related errors for every 53 nouns.

Thirdly, the regression line ($y_i$ = 0.3607$x$ + 3.6184) predicts an estimate of 22.7355 articles to be produced every 53 noun tokens. If we add the margin of error, the resulting interval predicts that 95% of the sampled population will produce a mean between 21.0541 and 24.4169 articles for every 53 nouns.

## 5. DISCUSSION

Although the aim of this study is it is not to analyze the types of article-related errors, or their possible motivations, a link can be established between our results and Ionin *et al.'s* (2004) Fluctuation Hypothesis. As can be seen in Figures 1–9, there is a great variance in terms of article-related errors, and the regression models – though they do not represent perfect fits – relate better to the observed number of errors than the mean line. This shows that the amount of article-related errors is hard to predict even in participants with a high degree of homogeneity, as is the present research.

With regard to our first research question, all the *t*-tests confirm statistically significant relations between the number of article-related errors and the number of

article tokens. This is also supported by the fact that none of the confidence intervals contain zero. This result was to be expected since more article tokens represent more linguistic contexts in which articles may occur.

Significant relations have also been found in relation to our second research question, that is, between the number of article-related errors and the number of noun tokens. Moreover, in this case the coefficients of determination ($r^2$) indicate good fits (over 50%). This seems to suggest that a piece of text which makes use of a higher number of noun phrases will require more articles by default and, as a consequence, the number of article-related errors can be expected to increase. That is why it is essential to analyze article-related errors in relation to the total number of noun phrases, i.e. article environments, and not just to the total amount of words per one hundred words. In order to deal with this issue, Testa (2019) suggests assigning an *error coefficient* (between 0 and 1) which results from dividing the total number of article-related by the total number of article environments. In this way, two separate learners who would make two errors in a total of four article environments, and fifteen errors in a total in thirty article environments would both share an error coefficient of 0.5. Conversely, a sentence like *Me gusta *uno vodka* ('me.DAT likes INDEF.MSG vodka') would mean an error coefficient of 2 (INDEF-for-DEF; M-for-F).

Finally, as for research question number 3, a significant relationship has been attested between the number of article tokens and the number of noun tokens. Moreover, the coefficient of determination ($r^2$) indicates a good fit in the B1 group. In other words, these are cases in which over 50% of the variance in article-related errors and article tokens can be explained by the total number of noun tokens. Nevertheless, the coefficient of determination with regard to the relationship between AT and NT should be taken with caution, because more frequent article usage can be rarely explained without resorting to a higher amount of linguistic contexts in which article take place, namely in noun phrases.

## 6. CONCLUDING REMARKS

The present study has drawn attention to the need to consider the number of article environments, i.e. article tokens in noun phrases, when analyzing article-related errors.

Statistically significant relations have been attested between these variables and they fit the data better than the mean average number of article-related errors an article tokens.

Moreover, the evidence with regard to LX English and LX Spanish article acquisition suggests that students lack knowledge as regards the semantic traits associated to article use. This calls for a deeper analysis of such traits, particularly at an intermediate level of learning given that the article system is one of those categories that does not seem to be acquirable by means of comprehensible input alone (cf. Pica 1985; Master 1994).

Finally, some key questions that have not been addressed in this study remain open. One aspect that calls for more in-depth treatment is the effectiveness of pedagogical interventions in terms of L2 article instruction (cf. Ekiert and Han 2016). For instance, it would be interesting to analyze whether a Spanish-oriented version of Master's binary framework is a helpful pedagogical tool when it comes to teaching the Spanish article system. Moreover, even though Master's (1997) distinction of two types of Ø article ('indefinite' $Ø_1$ vs. 'definite' $Ø_2$) seems to apply to Spanish in most cases, further research is required to show whether this distinction may accelerate the acquisition of the Spanish article system by [–ART] L1 learners.

## REFERENCES


Allan, Keith. 1986. *Linguistic Meaning*. London: Routledge.

Allen, Virginia F. 1983. *Techniques in Teaching Vocabulary*. Oxford: Oxford University Press.

Bailey, Natalie, Carolyn Madden and Stephen Krashen. 1974. Is there a natural sequence in adult second language learning? *Language Learning* 24/2: 235–243.

Birner, Betty and Gregory Ward. 1994. Uniqueness, familiarity, and the definite article in English. In Susanne Gahl, Andy Dolbey and Christopher Johnson eds. *Proceedings of the Twentieth Annual Meeting of the Berkeley Linguistics Society*. Berkeley, California: Berkeley Linguistics Society, 93–102.

Ekiert, Monika and ZhaoHong Han. 2016. L1-fraught difficulty: The case of L2 acquisition of English articles by Slavic speakers. In Rosa Alonso Alonso ed. *Crosslinguistic Influence in Second Language Acquisition*. Bristol: Multilingual Matters, 147–172.

Fernández Jódar, Raúl. 2006. *Análisis de Errores Léxicos, Morfosintácticos y Gráficos en la Lengua Escrita de los Aprendices Polacos de Español*. Poznań, Poland: Adam Mickiewicz University dissertation.

Fernández Jódar, Raúl. 2017. El artículo. In Wiaczesław Nowikow ed. *Gramática Contrastiva Español-polaco*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego, 353–377.

García Mayo, María del Pilar. 2008. The acquisition of four non-generic uses of the article the by Spanish EFL learners. *System* 36/4: 550–565.

George, Herbert. 1972. *Common Errors in Language Learning*. Rowley: Newbury House.

Hakuta, Kenji. 1976. A case study of a Japanese child learning English as a second language. *Language Learning* 26: 321–351.

Harb, Mustafa A. 2014. A closer look at the English article system: Internal and external sources of difficulty revisited. *International Journal of Linguistics* 6/4: 87–101.

Hawkins, John. 1991. On (in)definite articles: Implicatures and (un)grammaticality prediction. *Journal of Linguistics* 27/2: 405–442.

Hewson, John. 1972. *Article and Noun in English*. The Hague: Mouton.

Hidalgo, Andrea. 2015. *Estudio Contrastivo Español-chino: El Artículo Indefinido y su Tratamiento en los Manuales de Enseñanza de Español como Segunda Lengua*. Córdoba, Argentina: Universidad Nacional de Córdoba dissertation.

Ionin, Tania, Heejeong Ko and Kenneth Wexler. 2004. Article semantics in L2-acquisition: The role of specificity. *Language Acquisition* 12/1: 3–69.

Ionin, Tania, María L. Zubizarreta and Salvador Maldonado. 2008. Sources of linguistic knowledge in the second language acquisition of English articles. *Lingua* 118/4: 554–576.

Isabelli-Garcia, Christina and Rachel Slough. 2012. Acquisition of the non-generic definite article by Spanish learners of English as a foreign language. *OnOmázein* 25/1: 95–105.

Jaensch, Carol. 2008. L3 acquisition of articles in German by native Japanese speakers. In Roumyana Slabakova, Jason Rothman, Paula Kempchinsky and Elena Gavruseva eds. *Proceedings of the 9th Generative Approaches to Second Language Acquisition Conference*. Somerville, Massachusetts: Cascadilla Proceedings Project, 81–89.

Jenks, Peter. 2018. Articulated definiteness without articles. *Linguistic Inquiry* 49/3: 501–536.

Jiang, Nan, Eugenia Novokshanova, Kyoko Masuda and Xin Wang. 2011. Morphological congruency and the acquisition of L2 morphemes. *Language Learning* 61/3: 940–967.

Kharma, Nayef. 1981. Analysis of the errors committed by Arab university students in the use of the English definite/indefinite articles. *International Review of Applied Linguistics in Language Teaching* 19/4: 333–345.

Konieczna-Twardzikowa, Jadwiga. 1992. Caso y definitud en la lengua española desde la perspectiva polaca. *Estudios hispánicos* 2: 171–175.

Krashen, Stephen. 1982. *Principles and Practice in Second Language Acquisition*. London: Pergamon.

Krashen, Stephen and Tracy Terrell. 1983. *The Natural Approach: Language Acquisition in the Classroom*. Oxford: Pergamon Press.

Lapesa, Rafael. 2000 [1974]. Un, una como artículo indefinido en español. In Rafael Lapesa ed. *Estudios de Morfosintaxis Histórica del Español*. Madrid: Gredos, 477–487.

Lema, Rebeca. 2016. Las interferencias del español L2 en el estudio del gallego L2. *Itinerarios: Revista de Estudios Lingüísticos, Literarios, Históricos y Antropológicos* 23: 61–78.

Lin, Tzu Ju. 2003. *La Adquisición y el Uso del Artículo por Alumnos Chinos*. Alcalá de Henares, Spain: Universidad de Alcalá dissertation.

Little, David. 1994. Words and their properties: Arguments for a lexical approach to pedagogical grammar. In Terence Odlin ed. *Perspectives in Pedagogical Grammar*. Cambridge: Cambridge University Press, 99–122.

Lyons, Christopher. 1999. *Definiteness*. Cambridge: Cambridge University Press.

Master, Peter. 1986. *Measuring the Effect of Systematic Instruction in the English Article System.* Los Angeles, California: University of California (Unpublished Paper).

Master, Peter. 1987. *A Cross-linguistic Interlanguage Analysis of the Acquisition of the English Article System*. Los Angeles, California: University of California dissertation.

Master, Peter. 1990. Teaching the English articles as a binary system. *TESOL Quarterly* 24/2: 461–478.

Master, Peter. 1994. The effect of systematic instruction on learning the English article system. In Terence Odlin ed. *Perspectives on Pedagogical Grammar*. Cambridge: Cambridge University Press, 229–252.

Master, Peter. 1995. Consciousness raising and article pedagogy. In Diane Belcher and George Braine eds. *Academic Writing in a Second Language*. Norwood, NJ: Ablex, 183–204.

Master, Peter. 1997. The English article system: Acquisition, function, and pedagogy. *System* 25/2: 215–232.

Master, Peter. 2002. Information structure and English article pedagogy. *System* 30: 331–348.

Mizuno, Mitsuharu. 1999. Interlanguage analysis of the English article system: Some cognitive constraints facing the Japanese adult learners. *International Review of Applied Linguistics in Language Teaching* 37/2: 127–153.

Nowikow, Wiaczesław. 2017. Tiempos verbales. In Wiaczesław Nowikow ed. *Gramática Contrastiva Español-polaco*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego, 127–178.

Odlin, Terence. 1989. *Language Transfer*. Cambridge: Cambridge University Press.

Park, Sung B. 2006. *The Acquisition of Written English Articles by Korean Learners*. Carbondale, Illinois: Southern Illinois University dissertation.

Pica, Teresa. 1983. The article in American English: What the textbooks don't tell us. In Nessa Wolfson and Elliot Judd eds. *Sociolinguistics and Language Acquisition*. Rowley, Massachusetts: Newbury House Publishers, 222–233.

Pica, Teresa. 1985. The selective impact of classroom instruction on second language acquisition. *Applied Linguistics* 6/3: 214–222.

Pienemann, Manfred. 1998. *Language Processing and Second Language Development: Processability Theory*. Amsterdam: John Benjamins.

Ringbom, Håkan. 1987. *The Role of the First Language in Foreign Language Learning*. Clevedon: Multilingual Matters.

Ringbom, Håkan. 2011. Perceived redundancy or crosslinguistic influence? What L3 learners' material can tell us about the causes of errors. In Gessica De Angelis and Jean-Marc Dewaele eds. *New Trends in Crosslinguistic Influence and Multilingualism Research*. Bristol: Multilingual Matters, 19–24.

Ringbom, Håkan. 2016. Comprehension, learning and production of foreign languages: The role of transfer. In Rosa Alonso Alonso ed. *Crosslinguistic Influence in Second Language Acquisition*. Bristol: Multilingual Matters, 38–52.

Sabir, Mona H. 2015. *Explicit Instruction and Translation: A Generative View of the Acquisition of English Articles*. Leeds, United Kingdom: University of Leeds dissertation.

Schulz, Eckehard. 2004. *A Student Grammar of Modern Standard Arabic*. Cambridge: Cambridge University Press.
Schwarz, Florian. 2009. *Two Types of Definites in Natural Language*. Amherst, MA: University of Massachusetts dissertation.
Şekerci Arıbaş, Derya and Filiz Cele. 2019. Acquisition of articles in L2 and L3 English: The influence of L2 proficiency on positive transfer from L2 to L3. *Journal of Multilingual and Multicultural Development*. Advanced online publication .
Shen Jie. 2012. *El Artículo en la Enseñanza de ELE. Estudiantes de Origen Chino*. Barcelona, Spain: Universidad de Barcelona dissertation.
Sun, Ganzhao. 2016. The acquisition of English articles by second language learners: The sequence, differences, and difficulties. *SAGE Open* 6/1: 1–8.
Tarrés Chamorro, Iñaki. 2002. *El Uso del Artículo por Estudiantes Polacos de E/LE*. Barcelona, Spain: Universidad de Barcelona dissertation.
Testa, Martín. 2019. *Análisis de Variables Psicolingüísticas en la Interlengua de Alumnos Polacos de Español L3*. Warsaw, Poland: University of Warsaw dissertation.
Testa, Martín. In press. The acquisition of Spanish articles by L1 Polish students. *Neofilologia: Perspektywy Transdyscyplinarności*.
Trenkic, Danijela. 2002. Establishing the definiteness status of referents in dialogue (in languages with and without articles). *University of Cambridge Working Papers in English and Applied Linguistics* 7: 107–131.
VanPatten, Bill and Teresa Cadierno. 1993. Explicit instruction and input processing. *Studies in Second Language Acquisition* 15/2: 225–243.
Widdowson, Henry. 1988. Grammar, nonsense, and learning. In William Rutherford and Michael Sharwood Smith eds. *Grammar and Second Language Teaching*. New York: Newbury House, 146–155.
Yoo, Isaiah. 2009. The English definite article: What ESL/EFL grammars say and what corpus findings show. *Journal of English for Academic Purposes* 8/4: 267–278.

*Corresponding author*
Martin Testa
University of Warsaw
Institute of Iberian and Ibero-American Studies.
ul. Oboźna 8
00–332 Warsaw
Poland
e-mail: m.testa@uw.edu.pl

received: August 2019
accepted: October 2019

RiCL Research in Corpus Linguistics

# Designing the *Radiotelephony Plain English Corpus* (RTPEC): A specialized spoken English language corpus towards a description of aeronautical communications in non-routine situations

Malila Prado - Patricia Tosqui-Lucks
Universidade de São Paulo / Brazil

**Abstract** – Pilots and air traffic controllers need to undergo a specific English test in order to be granted a license for international operations. A language proficiency scale was developed to serve as a parameter to all aviation regulatory agencies throughout the world by targeting the language produced specifically by air traffic controllers and pilots in radio communications when non-routine situations (such as technical problems, bird strike, changes in weather, health problems on board, etc.) occur (ICAO 2010). However, there is a lack of empirical investigation which could shed light upon this particular register helping the users of the scale with its understanding. In an attempt to fill this gap, this paper outlines a compilation of the *Radiotelephony Plain English Corpus* (RPTEC), a spoken corpus of aeronautical communication consisting of transcriptions of exchanges between pilots and air traffic controllers in non-routine situations for research and pedagogical purposes. By presenting steps taken during the process, we intend to provide fellow researchers with data which may suit other purposes and yield further analyses, as well as enlighten similar investigations in the field of English for Specific Purposes.

**Keywords** – corpus design; spoken corpus; aviation language; English for Specific Purposes

## 1. INTRODUCTION[1]

Pilots and air traffic controllers (ATCOs) around the world are required to take a proficiency test which evaluates their ability to communicate in English on the radio in order to operate internationally. Such is the decision of the International Civil Aviation Organization (ICAO) after a series of accidents in which language was a contributory cause. By means of a document which prescribes procedures for both training in and testing of aviation English, ICAO (2010) determines that English proficiency of these

[1] This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

professionals is to be assessed according to six different criteria: pronunciation, structure, vocabulary, fluency, comprehension and interaction. Test takers need a minimum level 4 in all criteria to be granted the English proficiency license to operate internationally.

The lack of clarity in the description of aviation English *per se* is claimed to have offered different views of understanding (Alderson 2010). According to ICAO (2010), the scope of teaching and assessment is both standard aeronautical phraseology –a set of phrases and words that are fully described in official documents (Philps 1991)– and plain English, which is the language used when phraseology is not sufficient (ICAO 2010: x). The latter is a more spontaneous portion of language needed mostly when pilots and air traffic controllers face abnormal situations, that is, when something unexpected happens, since phraseology covers all routine situations of a flight. Despite being spontaneous, the plain English used in radio communications does not resemble informal spoken English, as this scope of language follows certain rules governed by aeronautical phraseology, such as being structurally concise and lexically restricted (Bieswanger 2016). Although routine communications should occur based on the standard aeronautical phraseology, the professionals often refer to plain English even in normal situations, particularly with pragmatic speech acts such as greeting or thanking one another (Lopez 2013).

Standard aeronautical phraseology (SAP) in English is documented by ICAO in different manuals (2001, 2007) as mandatory for international flights,[2] and it is usually learned and trained ever since pilots and ATCOs are in the initial process of their career. This scripted language focuses on precise and predetermined lexicon and reduced grammatical items with a view to eliminating ambiguity (Philps 1991; Moder 2013; Bieswargen 2016; Estival *et al*. 2016). The large number of aircraft under the control of the same professionals at the same time urges the adoption of specific call signs for each aircraft, which are airline and flight numbers (such as *GLO 323*, *Fastair 345*, *Speedbird 981*), instead of using personal pronouns (*I*, *we*, *you*). Likewise, the ATCO must be identified by the ground control facility (like Miami Center, Brasilia Control, GRU Tower) to avoid confusion. Communications should take place as illustrated in extract (1).

[2] According to ICAO documents, the standard phraseology must be used, whenever possible, in all phases of flight: for clearance and taxi, take-off and departure, read-back, climb, cruise and descent, approach and landing.

(1) **ATCO:** Fastair 345 when passing flight level 80 contact Alexander control 129.1

**Pilot:** When passing flight level 80 129.1 Fastair 345

The example above starts with an instruction given by the ATCO and is read back by the aircraft, whose call sign is *Fastair 345*. The use of call signs also prevents the ambiguity of deixis (see Garcia 2016). To further illustrate the difference between SAP and plain aviation English (see Bieswargen 2016), the latter being the object of this piece of research, example (2) presents the transcription of a real event.

(2) **Pilot:** Control / Aircraft 1010 / report //

**ATCO:** Go ahead //

**Pilot:** Aircraft 1010 / we have an indication of engine fire / procedures have been applied / we have no further uh confirming indications of engine fire / we are now recovering as a precautionary measure / I'll keep you posted // and uh souls on board uh seventy-nine passengers plus four crew / correction plus five crew //

**ATCO:** seventy-nine passengers plus five crew / roger / thank you // Aircraft 1010 / at this moment your position is uh eight miles on the right downwind // you are the number one for the approach now expect vectors for the ILS //

**Pilot:** We'll keep this heading for a while / and uh we will can perform a normal circuit for runway zero three / Aircraft 1010 //

**ATCO:** Aircraft 1010 // roger / can you confirm the engine with the problem? //

**Pilot:** It's engine number two / number two //

**ATCO:** Number two / roger //

Extract (2) illustrates the use of commands such as *report*, *go ahead*, *expect vectors for the ILS* and concise expressions such as *souls on board*, *correction* and *roger*, occurring alongside some elements, such as *I'll keep you posted*, *We'll keep this heading for a while*, *thank you*, as well as modal verbs, which can be easily found in a more spontaneous talk. These are examples of plain aviation English. The purpose of our study is not to describe how speakers use SAP, but the manner in which their communication happens when it exceeds the limits of this prescribed and documented language. ICAO's documents state that even in communications that are held when the unexpected happens, pilots and ATCOs should adhere to the principles of the SAP, which are objectivity, brevity and clarity.

There is considerable debate as to whether plain aviation English and SAP can be separated (Lopez 2013: 118). Such debate is based on the fact that both of them are interconnected, and it is difficult to establish the point at which SAP ends and plain

aviation English begins. The present research adopts the view that "standardized phraseology and plain aviation English can be categorized as two distinct specialized registers" (Bieswanger 2016: 74). By considering them as two concurrent registers, we selected parts of communications between pilots and controllers undergoing abnormal or emergency situations, in which SAP may be present, but most of the excerpt represents plain aviation English.

In order to understand how plain aviation English is used in this specific context, we have undertaken the compilation of a corpus of radio communications between pilots and ATCOs in non-routine situations, namely the *Radiotelephony Plain English Corpus* (RTPEC), which is presented in the following sections. The corpus was designed to meet both research and pedagogical purposes and, for this reason, it contemplates certain characteristics required for both areas.

This paper is structured as follows: firstly, Section 2 introduces the theoretical basis of corpus linguistics as an approach to the description of language. Then, Section 3 discusses some of the criteria involved in the corpus design, considering categories, speakers, texts, the transcription model and the current amount of data. The paper concludes with some suggestions for future research.

## 2. CORPUS LINGUISTICS

Corpus linguistics is a research method which employs corpora for data extraction. A linguistic corpus is a bank of texts stored in computers, which allow for a (semi-) automatic extraction of data by using statistical analysis. Spoken texts must be transcribed as the computer software commonly used in such investigations is fed by the written word.[3]

Studies carried out with this methodology very often highlight evidence usually not perceived by the naked eye (Sinclair 2004). For this reason, a number of features have been brought out by such studies and have particularly favored fields related to dictionary-making, translation studies and, to a lesser extent, coursebooks, to mention but a few. A perspective underlying studies based on corpus linguistics is that language is stored

[3] We are currently not aware of any tool that recognizes the spoken word without its corresponding written form.

cognitively and retrieved in chunks, rather than in single words, and built through social use, therefore conventionalized (Wray and Perkins 2000; Wood 2006).

Conventionalized language is revealed by the high frequencies with which certain elements occur in a given linguistic community. The more often they occur, the more conventionalized they are. Such frequencies can only be identified through corpora, and their relevance –particularly when it comes to corpus design– relies on the context from which the language was taken. Therefore, the origin of the texts along with their context of production is of utmost importance. The texts must represent the use of the language in the context of production being investigated. Some writers advocate for as faithful a representation of the spoken language as possible, even though it is questionable whether the written form can actually cover all the non-verbal aspects present in real interaction (Haberland 2010).

The corpus should encompass the linguistic community from which the texts are extracted. In recent years corpora have become larger as computer capabilities have increased (the 560-million-word *American Corpus of Contemporary English*[4] is an example). However, some authors defend the use of smaller corpora for specific context investigations, such as studies in the context of Language for Specific Purposes (Gavioli 2005) or Pragmatics (Vaughan and Clancy 2013), as few occurrences may be enough to represent the language characteristics of these groups of speakers. That is the case of aviation English, due to the fact that the language is concise, the vocabulary is restricted and the context is very limited. Aviation English itself has witnessed studies in corpora mainly with the objective of describing language use; some examples can be found in Sarmento (2008), Bocorny (2008), Lopez (2013) and Tosqui-Lucks (2018).[5]

## 3. THE COMPILATION PROCESS

It is commonly stated that spoken corpora are challenging for many different reasons: the compilation itself, copyright issues, the choice for spontaneous or guided language, the quality of video and audio equipment and the transcriptions (Adolphs and Knight 2010).

[4] https://corpus.byu.edu/coca/

[5] Tosqui-Lucks (2018) described 16 studies, from different parts of the world, from the 1990s to the present date, all dealing with aviation English and Corpus Linguistics.

These issues were faced in this piece of research as well. In this section, we elaborate on each of the choices made throughout the process.

### *3.1. The criteria*

According to Bieswanger (2016: 74), "[w]hile standardised phraseology is concerned with the fairly restricted aspects of routine air traffic control issues, plain Aviation English covers a broader range of topics in non-routine situations, such as emergencies as well as other unusual or unexpected contexts." Considering this, the first decision regarding our corpus was to select audios in which operational problems occur, these being of human, weather or mechanical nature, insofar as the oral communication happens between ATCOs and pilots. Our first endeavor was the compilation of transcripts of accident reports; however, we soon realized that these transcripts consider mostly the communication happening inside the flight deck, thus not corresponding to the scope of language being studied. Moreover, as prosody might have relevance in the analysis, all transcriptions should necessarily have their audio files. Finally, the selection of the events should be focused on solvable problems whenever possible, as the corpus texts are intended for students of aviation English.

To this end, we resorted to a well-known website in the area of aviation English called Live Air Traffic,[6] which stores communications held in different parts of the world, especially the United States. Events representing non-routine situations are rare, and the overwhelming amount of data being stored 24 hours a day at different airports worldwide hindered our search. One of the sections of this website solves the issue of finding communications with technical problems by providing a bank of audios of different kinds. Our goal was then communications of non-routine matters, but they still needed to be carefully selected. After collecting the audios, we verified their source in different websites: newspapers, TV news, accident/incident databases and *The Aviation Herald*.[7]

The time period of the communications selected was also taken into account. As there was no control over the identification of the speakers, we chose to collect files of events which happened from 2008 onwards, the year when the ICAO proficiency

[6] https://www.liveatc.net
[7] A web 2.0 model site built with the participation of users and members (https://www.avHerald.com).

requirements referred to in Section 1 of this paper should be implemented by all state members.

During the process of beta-testing of the data compiled, we came across some problems. The first problem was related to the two most frequent content words in the corpus: *runway* and *engine*. The word *runway* corroborated aviation investigations attesting that most accidents happen on or near the runway.[8] However, *engine* could mean an excess of audio files related only to engine problems, which was verified by searching for the main abnormal situation represented in each of the audios. We then noticed that there were only events with engine malfunctions, which allowed for few potential findings related to other terms. Thus, in order to increase the representativeness of the corpus, we tried to widen the variety of problems by choosing sources of events in different data: annual accident and incident reports, manufacturers' statistics and governments' reports; however, occurrences vary depending on the countries (some hold better safety records than others), manufacturers (as their aircraft systems are different) and time (aviation has become safer over time).

Any of these choices would be random. We thus decided to use a document entitled *Taxonomy of Occurrences* (ICAO 2006), composed of a script of categories which standardize accident and incident reports internationally as a means of diffusing information. There were other possible choices, but adhering to an official text which needs to be used when abnormal situations occur seemed to be a better way of approaching our selection standards. This document enumerates 33 different categories of occurrences in aviation, from minor to major incidents or accidents,[9] such as bird strike, fire/smoke, fuel related, collision with obstacles, system/component failure and malfunction, among others. Based on this taxonomy, the initial idea was to compile at least three audios for each category.

By assuring that the audios met the criteria listed in the taxonomy, we started to obtain better results in terms of lexical range. The corpus was finalized when it reached 12 hours of audio material transcribed into 110,737 words, in a total of 130 texts about 31 different occurrence categories. The transcriptions are monolingual and non-annotated, as we intended to start from a corpus-driven study (Tognini-Bonelli 2001).

[8] See https://www.boeing.com/news/techissues/pdf/statsum.pdf (20 December, 2018).
[9] Broadly speaking, incidents are minor occurrences whose consequences are avoided, and accidents refer to damage to aircraft or injuries to people or even fatalities.

*3.2. Corpus design*

In our approach we argue in favor of aviation English as a lingua franca (Estival *et al*. 2016), as English is the language of aviation, and we assume that native and non-native speakers in aviation use the language in a similar manner. Previous studies have also shown that even native speakers of the language need to adhere to the aviation English standards, as aviation English is not a register acquired in daily life events (Bieswanger 2016). The objective is to compile a corpus of operational users of a vocational language, regardless of these users' mother tongue.

In a latter phase of the research, we suspected that the corpus was biased towards the North-American culture rather than representing patterns of an international community, since most audios collected were originally held in the United States. There was a need to implement the corpus with a more international variety of speakers. Nevertheless, with the continuous growth of international flights, it has become easier to find exchanges among different nationalities. Still, it is impossible to determine if the speakers are native speakers of English, bilinguals or even non-native speakers. Some airline companies are widely known for the high presence of international pilots in their payroll, especially those in the Middle East and Asia. Besides, the English proficiency of the transactions compiled has clearly improved in the late years (see Prado and Tosqui-Lucks 2017), which might be a result of the implementation of the proficiency requirements. This enhancement seems to point out that aviation English is becoming more widely used by an international community.

Seeking a broader representation of the professional community, we opted to add at least one more audio file to each of the categories suggested by the *Taxonomy of Occurrences* (ICAO 2006), inasmuch as it contained exchanges between international traffic, that is, a foreign aircraft in a particular airport or airspace. Whenever an airline company is considered foreign to a certain airport or airspace, radio communications are held in English even if the pilots and ATCOs share the same mother tongue. Besides, native speakers of English must allegedly accommodate their English to an international speech community (Estival *et al.* 2016) provided they are not in their local environment.

*3.3. The transcription*

Restricted to the scope of problems, the transcription starts when the problem is first mentioned, and finishes when the problem is either solved or handed out to another professional (i.e. a fire fighter or maintenance technician). For this reason, the size of the texts differs (some are less than a minute long and others are longer than 15 minutes).

By using the *Taxonomy of Occurrences* as a basis, we included the audio file names, time of each audio and the final count of words to organize the corpus. Each occurrence has a number of four to six audios to embrace as large a number of lexical items as possible. We also inform about the duration of each file, followed by the number of words transcribed and the final count of words for each category.

All the corpus texts were manually transcribed, as the sound quality, number of speakers and different accents hinder the use of any speech recognition software, which needs to be trained to a specific voice to enable the automatic transcription. The minimum number of participants in radio communications is two, but there is often a larger number when other aircraft are involved in the event. To assure control of features of the transcription which needed to be strictly connected to the context of production, the decision was to concentrate the transcription on one linguist researcher. The transcriptions were then reviewed by linguists specialized in aviation English and subject-matter experts (SMEs), namely pilots or ATCOs with international experience. Whenever possible, pilots and ATCOs from the region of a specific occurrence were requested to review the transcription, as they are supposedly more familiar with the waypoints, departure and arrival points, airport names and call signs present in the audio files.

Of great assistance during the transcription was the software *SoundScriber,*[10] used by the University of Michigan in corpus compilation. It allows the transcriber to control the audio file without leaving the text processor. By means of shortcuts in the keyboard, the transcriber saves time by playing, pausing and rewinding the audio file; the loop, one of the most resourceful tools, replays a time segment continuously, moving ahead little by little, adjusted by the transcriber according to his/her typing speed (see Figure 1).

[10] See http://www-personal.umich.edu/~ebreck/code/sscriber/

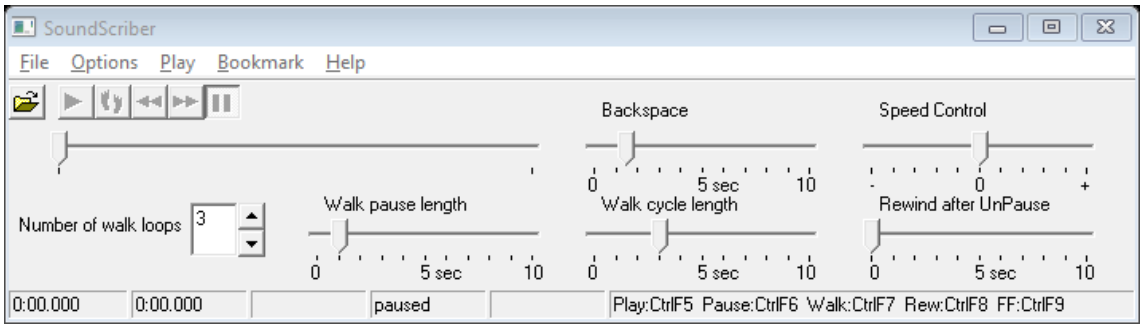


Figure 1: *SoundScriber* main screen

The spoken language is not governed by the same rules as the written language. Pauses and hesitations, for example, do not correspond to the syntactic rules of comma usage, and neither do breaks or false starts. The attempt to transform spoken language in a representation which can allow for linguistic investigations is a political issue (Haberland 2010: 62) for two main reasons: (i) we take the product from its source of production and it immediately becomes abstract; and (ii) a transcription is not the event itself, but a representation of the event. Still, the transcription needs to be linked to its source of production (such as speakers, event, place, time), as a means of preserving the analysis and readability of the texts.

During the first transcriptions, we faced problems as illustrated in extract (3).

(3) **ATCO:** <aircraft call sign> while you're waiting uh just the reason for the uh wait is the tail strike may have been in the intersection just waiting for the inspection to ensure that the runway is serviceable.

The utterance in (3) is lengthy, and it is difficult to recognize where the pauses are. Adding commas would not represent the particularity of such an utterance, and it is also problematic to try to identify where punctuation marks may fit. In order to sort out this problem, we adopted a prosodic and pragmatic framework called the 'Language into Act Theory' (L-AcT) (Cresti 2000). Based on how the utterance was produced, the prosodic breaks are identified by adding slashes separating the tone units, as can be seen in extract (4).

(4) **ATCO:** <aircraft call sign> while you're waiting uh / just the reason for the uh wait is the tail strike may have been in the intersection // just waiting for the inspection to ensure that the runway is serviceable //

One slash (/) represents a break in the flow within the utterance; two slashes (//) indicate the end of the utterance, identified in the intonation. However, the end of the utterance does not necessarily mean the end of the turn, which might be held by the speaker. Because it is an oral production, this language is better approached as utterances rather

than sentences, which was facilitated by using the L-AcT. This choice would favor an investigation of the so-called conversation or spoken grammar (Rühlemann 2008), claimed to be different from the written form.

As the corpus was also built for pedagogic purposes to students of aviation English, the texts were preserved as clean of metalinguistic information as possible. Besides, the texts should be readable by research tools, allowing for a more clear-cut investigation. Separating tone units by lines, for example, would hamper the reading of the concordance lines, as the utterances would appear altogether. The slashes enable manual extraction in the sense that the units are visibly separated, even with segments corresponding to false starts or corrections, for example.

The first part of the text is the context of production, which contains the source of the audio file and any other database where a report of the event might be found, along with keywords to identify the problem and the professionals who have proof-read the transcription. This information goes between the tags <header> and </header>, identifying beginning and end of the setting (see Appendix 1 for a sample text).

Other tags used in the transcriptions to mark metalinguistic information include the following:

- <unreadable> refers to parts which none of the transcribers could understand, either because they were poorly pronounced or were waypoints specific of the location.
- <false start> identifies utterances that were initiated and then rephrased, and <break> identifies a break in the flow. They help the reader in the organization of certain disfluency features one might find in spoken language.
- <blocked transmission> means there was an overlap, which produces noise when two speakers hold the radio push-to-talk button at the same time.
- <pause> is used for silent pauses.
- <noreply> is introduced whenever one of the participants does not provide any feedback for the input previously given. This silence is part of the conversation, as it elicits the previous speaker to use strategies so as to finally obtain the response needed.

Some verbal aspects were also transcribed, even though they are not words *per se*, such as disfluencies: *uh* (for filled pauses), *uhh* (for longer filled pauses) and other noises. Numbers were transcribed in full for investigations of standard vs. non-standard phraseology, and words such as *okay* and *alright* were transcribed as such in order to avoid confusions with the letters *O* and *K* (for *okay*) and with the words *all* and *right*.

We opted to remove call signs of aircraft and ground stations when we printed out the transcriptions in paper publications or classroom handouts for the sake of preserving the identity.

To assess the representativeness of the corpus, there were constant measures of the type-token ratio (TTR) of the corpus, that is, the amount of running words divided by the number of types. Token refers to the total occurrences of a word, whereas type refers to the number of different words. This calculation reveals how dense or complex a corpus is; the higher the TTR, the richer the corpus. Corpus Linguistics tools such as *Wordsmith Tools* (Scott 2016) can quickly assist the researcher with calculating such complexity. The corpus of this study has shown to be of low lexical density, as its TTR oscillated between 6% and 3.68% in its first 20 transcriptions. Summing up, with the current 110,737 words, the final TTR has reached 2.9%; such a low density indicates that the corpus has highly repetitive patterns.

## 4. CONCLUSIONS AND SUGGESTIONS FOR FUTURE RESEARCH

The description presented in this paper aimed at providing a new contribution which brings together research projects with a shared interest.

The potential uses of the corpus are manifold, among others:

(i) Investigations on the use of the language by specific local communities;

(ii) a focus on the negotiation of meaning or other strategies that pilots and ATCOs use when trying to solve problems collaboratively;

(iii) an examination of the use of language by native and non-native speakers;

(iv) an analysis of aspects which are more intrinsic to language or metalinguistics such as a study of the linguistic areas prescribed in the ICAO scale, among others.

The present corpus has already rendered an investigation of the most frequent lexico-grammar patterns, aiming at a comparison with the structure and vocabulary areas listed in the ICAO scale. It has also been used for the analysis of fluency and interaction (other two linguistic areas prescribed by the ICAO scale). However, such studies are not exhaustive, and so far have only presented language to be prioritized in the aviation English classroom.

The corpus can be enhanced as an ongoing process to bring about investigations of different features. In our specific case, we seek a description that yields an empirical understanding of aviation English to support curriculum design and material development in order to provide tools for the field of teaching aviation English. Nevertheless, we do not exclude the benefits which this corpus can bring to the understanding and elaboration of tests, as it can provide test designers with materials to try to approach meaningful interaction with the test candidates.

Nonetheless, there are some caveats which need to be addressed in future research; one of them is related to the pronunciation features which are not yet linked to the written data. Despite the existence of software that enables the paralleling of the audio and written data, it is far time-consuming, and at present we lack the time or funding to do so. Still, a possible alignment considering utterances would generate broader understanding of pronunciation matters.

It should be emphasized that this corpus represents a very specific scope of an already specific field: aviation English. This corpus could be incorporated into a larger project containing other corpora collected from other parts of the world or even involving other aviation professionals.

The investigations with which corpora equip the researcher and material designer offer the opportunity to bring authentic language to the classroom, supplying the language professional with adequate tools to teach from meaningful input and, as such, promote better output from the students. We hope to contribute to the field of aviation English with this research.

## REFERENCES

Adolphs, Svenja and Dawn Knight. 2010. Building a spoken corpus: What are the basics? In Anne O'Keeffe and Michael McCarthy eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 38–52.

Alderson, Charles. 2010. A survey of aviation English tests. *Language Testing* 27/1: 51–72.

Bieswanger, Markus. 2016. Aviation English: Two distinct specialised registers? In Christoph Schubert and Christina Sanchez-Stockhammer eds. *Variational Text Linguistics: Revisiting Register in English*. Berlin: Mouton de Gruyter, 67–85.

Bocorny, Ana. 2008. *Descrição das Unidades Especializadas Poliléxicas Nominais no Âmbito da Aviação: Subsídios para o Ensino de Inglês para Fins Específicos (ESP).* Porto Alegre: Universidade Federal do Rio Grande do Sul dissertation.

Cresti, Emanuela. 2000. *Corpus di Italiano Parlato (Vol. I).* Firenze: Accademia della Crusca.

Estival, Dominique, Candace Farris and Brett Molesworth. 2016. *Aviation English: A Lingua Franca for Pilots and Air Traffic Controllers*. London: Routledge.

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

Garcia, Angela. 2016. Air traffic communications in routine and emergency contexts: A case study of Flight 1549 'miracle on the Hudson'. *Journal of Pragmatics* 106: 57–71.

Haberland, Hartmut. 2010. Pragmatics as a component vs. pragmatics as a perspective of linguistics. *Studies in Pragmatics* 12: 54–68.

International Civil Aviation Organization (ICAO). 2001. *Annex 10 to the Convention on International Civil Aviation: Aeronautical Telecommunications.* Montreal: International Civil Aviation Organization.

International Civil Aviation Organization (ICAO). 2006. *Aviation Occurrence Categories: Definitions and Usage Notes.* Montreal: International Civil Aviation Organization.

International Civil Aviation Organization (ICAO). 2007. *Manual of Radiotelephony DOC 9432-AN/925.* Montreal: International Civil Aviation Organization.

International Civil Aviation Organization (ICAO). 2010. *Manual of Implementation of the Language Proficiency Requirements (DOC9835-AN/453)* (second edition). Montreal: International Civil Aviation Organization.

Lopez, Stephanie. 2013. *Norme(s) et Usage(s) Langagiers: Le Cas des Communications Pilote-contrôleur en Anglais.* Toulouse: Université Toulouse le Mirail dissertation.

Moder, C. Lynn. 2013. Aviation English. In Brian Paltridge and Sue Starfield eds. *The Handbook of English for Specific Purposes*. West Sussex: Wiley-Blackwell, 227–242.

Philps, Dennis. 1991. Linguistic security in the syntactic structures of air traffic control English. *English World-Wide* 12/1: 103–124.

Prado, Malila C. A. and Patricia Tosqui-Lucks. 2017. Are the LPRs focusing on real life communications issues? *International Civil Aviation English Association*. Dubrovnik: Embry-Riddle Scholarly Commons, 1–20. https://commons.erau.edu/cgi/viewcontent.cgi?article=1027&context=icaea-workshop (22 October, 2019.)

Rühlemann, Christoph. 2008. A register approach to teaching conversation: Farewell to standard English? *Applied Linguistics* 29/4: 672–693.

Sarmento, Simone. 2008. *O Uso dos Verbos Modais em Manuais de Aviação em Inglês: Um Estudo Baseado em Corpus.* Porto Alegre: Universidade Federal do Rio Grande do Sul dissertation.

Scott, Michael. 2016. *Wordsmith Tools version 7*. Liverpool: Lexical Analysis Software.

Sinclair, John. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work.* Amsterdam: John Benjamins.
Tosqui-Lucks, Patricia. 2018. Aplicações de corpora no ensino e na avaliação de inglês aeronáutico: Estado da arte, reflexões, direcionamentos. In Matilde Scaramucci, Patricia Tosqui-Lucks and Silvia M. Damião eds. *Pesquisas sobre Inglês Aeronáutico no Brasil*. Campinas: Pontes, 89–114.
Vaughan, Elaine Claire and Brian Clancy. 2013. Small corpora and pragmatics. In Jesús Romero-Trillo ed. *Yearbook of Corpus Linguistics and Pragmatics*. Dordrecht: Springer, 53–73.
Wood, David. 2006. Uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency. *Canadian Modern Language Review* 63/1:13–33.
Wray, Alison and Michael R. Perkins. 2000. The functions of formulaic language. *Language and Communication* 20: 1–28.

*Corresponding author*
Malila Prado
Av. Prof. Luciano Gualberto, 403
3º andar – sala 14
Cidade Universitária- Butantã
05508-010 São Paulo - Brasil
e-mail: malilaprado@usp.br

received: August 2019
accepted: October 2019

**Appendix 1: Sample text**

<header>

Ukraine Airlines B737 x KLM F-70 at Brussels

Sep 11 2016

http://avherald.com/h?article=49de7586&opt=0

Venna

</header>

One two six six two five / Aircraft eight three kilo echo / good evening //

Aircraft International one four six / stop immediately / I say again / stop immediately //

Aircraft International one four six / hold position // Aircraft one seven two five / go around / I say again / go around / immediate right turn heading zero one zero //

Going around on heading zero one zero / Aircraft one seven two five //

Aircraft one seven two five / climb three thousand feet //

Say again / one seven two five? //

Aircraft one seven two five / climb three thousand feet //

Climbing three thousand / one seven two five //

Aircraft seven eight Quebec tango / line up two five right //

Line up two five right / Aircraft seven eight Quebec tango //

Aircraft eight echo x-ray / established ILS two five left //

Aircraft nine correction Aircraft eight echo x-ray / City Tower / hello / continue approach two five left / number 2 / wind one five zero degrees four knots //

Continue / eight echo x-ray //

Aircraft one seven two five / contact arrival again one one eight decimal two five //

One eight two five / Aircraft one seven two five/ speak to you in a minute or two //

Aircraft seven eight Quebec tango / wind one two zero degrees two knots / two five right / cleared for take-off //

Cleared for take-off two five right / Aircraft seven eight Quebec tango //

Aircraft International one four six / contact departure one two six decimal six two five / goodbye //

One two six six two five departure / thank you / Aircraft International one four six //

RiCL Research in Corpus Linguistics

# Changes in argument structure in Early Modern English with special reference to verbs of DESIRE: A case study of *lust*

Noelia Castro-Chao
University of Santiago de Compostela / Spain

**Abstract** – In Old and Middle English, several verbs of DESIRE could be found in impersonal constructions, a type of morphosyntactic pattern which lacks a subject marked for the nominative case controlling verbal agreement. The impersonal construction began to decrease in frequency between 1400 and 1500 (van der Gaaf 1904; Allen 1995), a development which has been recently investigated from the perspective of the interaction between impersonal verbs and constructional meaning by Trousdale (2008), Möhlig-Falke (2012) and Miura (2015). This paper is concerned specifically with the impersonal verb *lust* (< ME *lusten*) as a representative of Levin's (1993) class of verbs of DESIRE, some of which developed into prepositional verbs in Present-day English. The main aim here is to explore the changes undergone by *lust* during the two centuries after it ceases to appear in impersonal constructions, as well as to reflect upon some of the possible motivations for such changes. The data are retrieved from *Early English Books Online Corpus 1.0*, a 525-million-word corpus, and the examples are analysed manually paying attention to the range of complementation patterns documented in Early Modern English (1500–1700).

**Keywords** – argument structure; corpus linguistics; Early Modern English; impersonal construction; impersonal verb; verbs of DESIRE

## 1. INTRODUCTION[1]

The present paper explores the historical development of the verb *lust* in Early Modern English (1500–1700; henceforth EModE), a member of the class of verbs of DESIRE as defined in Levin (1993: 194–195) in her discussion of Present-day English (henceforth PDE) patterns of verb alternation. Verbs of DESIRE include, among others, formerly

[1] For generous financial support, I am grateful to the following institutions: the Spanish Ministry of Education (grant FPU2014/03208), the European Regional Development Fund, the Spanish Ministry of Science, Innovation and Universities (grant FFI2017-86884-P) and the Regional Government of Galicia (Directorate General for Scientific and Technological Promotion, grants ED431D 2017/09 and ED431B 2017/12). Thanks are also due to Teresa Fanego and Nuria Yáñez-Bouza for their valuable feedback on an earlier version of this paper, to Tamara Bouso-Rivas for her helpful comments and suggestions (personal communication) and, last but not least, to the anonymous reviewers and the editors of *Research in Corpus Linguistics* for their time and consideration.

impersonal verbs such as *hunger*, *long*, *lust* or *thirst* (Levin 1993: 194–195), which in PDE have developed prepositional objects (e.g. *pregnant women* ***lusting*** *for pickles and ice cream*. *Lexico's Dictionary* s.v. *lust* v.). Impersonal verbs of DESIRE have been found to alternate between impersonal and personal use in Old (500–1100) and/or Middle English (1100–1500; henceforth OE and ME, respectively), as illustrated in examples (1) and (2) below from the *Middle English Dictionary* (MED).

(1) He was for-hungred and **lust** to eten.
he-SUBJ was starved and lusted to eat
'He was starved and desired to eat'
(MED [c1390] Chart.Abbey HG [LdMisc 210] 353)

(2) Me **lust** no lenger lyue
I-OBJ wish no longer live
'I do not wish to live any longer'
(MED [a1400] *Cursor* [Trin-C R.3.8])

Example (1) illustrates a personal construction with a grammatical subject, *he*. Example (2), by contrast, illustrates an impersonal construction in which a grammatical subject is missing. In both examples, the first argument encodes the semantic role of EXPERIENCER, which represents the "animate being inwardly affected by an event or characterized by a state" (Traugott 1972: 34; see also Möhlig-Falke 2012: 31, fn. 12; Miura 2015: 6). According to McCawley (1976: 194), in impersonal constructions the EXPERIENCER may be said to denote a human being who is "unvolitionally involved in the state of affairs" expressed by the verb and who cannot, therefore, be conceptualised as the causer of the event or process.

The second argument encodes the semantic role of CAUSE, which in (1) is syntactically realised by a *to*-infinitive clause, *to eten*, and in (2) by a bare infinitive, *no lenger lyue*. Semantically, the CAUSE argument represents "something from which the experience emanates or by which the experience is effected" (Fischer and van der Leek 1983: 346).

In English, the impersonal construction is known to have started to disappear between 1400 and 1500 (van der Gaaf 1904: 142; Allen 1995: 441–442), but marginal impersonal instances are recorded until about two centuries later[2] (see Visser 1963: §43; Möhlig-Falke 2012: 14–15). Thus, the EModE period, with which this study is

[2] As in the following example extracted from the *Oxford English Dictionary* (OED): a1556. *Let hym come when hym lust*. OED s.v. *lust*, v. †2.

specifically concerned, is of interest from a historical point of view since impersonal verbs were in the process of readjusting their argument structure to the new possibilities of the grammatical system of English. A wide variety of factors have been claimed to affect the loss of impersonal patterns, giving rise to an extensive literature on the topic, which includes classical works in historical linguistics dating back to the early twentieth century (e.g. van der Gaaf 1904; Jespersen 1961: 208ff), followed by publications like McCawley (1976), Elmer (1981), Fischer and van der Leek (1983), Allen (1986, 1995) and more recently Trousdale (2008), Möhlig-Falke (2012) and Miura (2015), among many others.

After the loss of impersonal patterns, impersonal verbs developed a very idiosyncratic range of syntactic uses, some of which co-existed with impersonal patterns already in OE, as has been shown in previous work (e.g. Fischer and van der Leek 1983; Allen 1995: 286–287). It thus appears that

> the loss of impersonal patterns proceeded over the respective verbs in a very gradual and seemingly unsystematic manner, in that individual verbs developed in different syntactic ways. (Möhlig-Falke 2012: 3–4)

The overall aim of the present study is to elucidate the path of development followed by formerly impersonal verbs of DESIRE, focusing initially on the verb *lust* as a case study. Future work will address the development of the other impersonal members of the class (*hunger*, *long* and *thirst*), so as to obtain a complete picture of this semantic class, which, as pointed out by Miura (2015: 244), has received little or no attention in the literature to date. The verb *lust* in particular has been selected as the object of study here because the meaning most commonly expressed by this verb until the mid-sixteenth century is 'to desire' (e.g. a1425. *No creature shal luste* [i.e. desire] *play* [...]. OED s.v. *lust*, v. †3. a.; see also van der Gaaf 1904: 74–75). It is not until the late seventeenth century that the specialised sense 'to have a carnal desire' gains ground and survives up to the present day, though as a low-frequency usage (e.g. *He really* ***lusted*** *after me in those days*. *Lexico's Dictionary* s.v. *lust* verb).

As regards the specific objectives of this paper, they can be summarised as follows: 1) to determine the time when the verb *lust* exactly ceased to be recorded in impersonal constructions; 2) to provide a diachronic overview of the personal morphosyntactic patterns that came to replace impersonal constructions with this verb in the EModE period; 3) to describe the syntactic and semantic properties of the arguments of *lust*; 4) to

reflect upon some of the factors which have been claimed to affect the loss of impersonal patterns in the history of English.

Section 2 offers an overview of the development of impersonal constructions in earlier English and the main hypotheses put forward in the literature about their disappearance. Section 3 is concerned with the syntactic and semantic properties of the class of verbs of DESIRE. Section 4 outlines the data sources and methodology employed in the study, while Section 5 looks at the origin and development of *lust* as well as the range of complementation patterns documented with this verb based on the entries of the OED, the MED and previous studies, looking at both impersonal and personal patterns. Section 6 presents and discusses the data on *lust* retrieved from the *Early English Books Online Corpus 1.0* (1473–1700; henceforth EEBOCorp 1.0). Section 7 summarises the main findings and conclusions to be drawn from the study.

## 2. THE DEVELOPMENT AND LOSS OF IMPERSONAL CONSTRUCTIONS

Before we delve into the question of impersonal constructions, a few comments are in order regarding the use of the terms 'impersonal' and 'personal'. In my use of the term 'impersonal', I will follow Fischer and van der Leek (1983: 347) and Möhlig-Falke (2012: 6) in treating as impersonal those morphosyntactic patterns which lack a grammatical subject controlling verbal agreement. The term 'personal', conversely, will be applied to patterns which involve a grammatical subject controlling verbal agreement. For a discussion of the terminological and conceptual maze surrounding impersonal constructions, see Méndez Naya and López Couso (1997).

Various hypotheses have been put forward to try to explain the causes which may have led to the disappearance of impersonal constructions. Most notably, Jespersen's (1961) reanalysis hypothesis has dominated the discussion throughout the twentieth century and, in spite of the criticisms it has received, it has remained a major topic of discussion in the works of Fischer and van der Leek (1983) and Allen (1986, 1995), among many others. According to Jespersen (1961: 208–210), the EXPERIENCER argument in impersonal expressions underwent a process of reanalysis as a result of the syncretism of forms brought about by the simplification of the case system. Examples (3a)–(3d) below represent the hypothetical stages postulated by Jespersen (1961: 209) in order to account for the changes involved, which include, first, the richly inflected sentence in

(3a), representative of OE; second, the syncretism of case forms in the nominative and the dative represented in (3b), corresponding to early ME; and, third, the structural ambiguity in (3c), which eventually led to a confusion about which constituent functioned as subject and object of the clause. This structural ambiguity arose in OVS patterns with two NPs, like (3c), probably representative of late ME, and it eventually cancelled the possibility to place the oblique EXPERIENCER before the verb once word order became rigidified (Fischer and van der Leek 1983: 338–339). Thus, by the EModE period the EXPERIENCER was reanalysed as a subject, as represented in example (3d), with the morphologically marked pronoun *he* as the unambiguous subject of the clause.

(3a) þam cynge **licodon** peran
the king-DAT/SG liked-PL pears-NOM/PL
'pears pleased the king'

(3b) the king **liceden** peares
the king-SG liked-PL pears-PL
'pears pleased the king'

(3c) the king **liked** pears
'pears pleased the king/the king liked pears'

(3d) he liked pears

According to Jespersen (1961: 208), the "natural" outcome was for the EXPERIENCER to be reanalysed as subject, mainly due to "the greater interest taken in persons than in things, which caused the name of the person to be placed before the verb." Aside from this psychological explanation, Jespersen's account bears on the deep morphosyntactic transformations which the English language underwent during OE and ME, namely the simplification of the case system, which has been dated in the twelfth and thirteenth centuries (Allen 1995: 213, 441) and the rigidification of word order, dated in the mid-fifteenth century (Fischer *et al.* 2000: 162; see also Möhlig-Falke 2012: 19, 216).

The reanalysis hypothesis, however, has been challenged on the basis of empirical data showing that impersonal patterns remained productive even after these changes were becoming complete (see e.g. Fischer and van der Leek 1983; Allen 1986, 1995). Allen (1986) in particular notes that Jespersen's claim cannot be upheld if we take into account that sentences such as (3a)–(3c) with two nominal NPs are actually highly infrequent in OE and ME for the impersonal verb *like* (see Allen 1986: 378). The reason for this is that it is not likely that the loss of case marking played a role if we bear in mind that case

distinctions on pronouns remained clear in the majority of cases, so that formal ambiguity did not arise as a rule (see also McCawley 1976: 201–202; Fischer and van der Leek 1983: 339, 346ff).[3]

Even though the traditional account assumes that it was the EXPERIENCER argument rather than the CAUSE which was reanalysed as subject, several studies have pointed out that impersonal verbs in fact developed along various distinct syntactic paths (Fischer and van der Leek 1983: 365–366; Möhlig-Falke 2012: 217). For the purposes of this study, two of these paths are described in the paragraphs that follow (for a full account see e.g. Möhlig-Falke 2012: 217–218).

> **Path 1**: The EXPERIENCER argument is interpreted as subject and the CAUSE argument, if expressed, is encoded as object. This path corresponds to so-called EXPERIENCER-subject constructions (Fischer and van der Leek 1983: 352–354) and it is the most common path of change of impersonal verbs (Möhlig-Falke 2012: 218). It is the path followed by *hunger*, *like*, *need* or *thirst* (e.g. *She likes money*. Fischer and van der Leek 1983: 363).
>
> **Path 2**: The EXPERIENCER is interpreted as object and the CAUSE is encoded as the subject of the clause. This path corresponds to so-called CAUSE-subject constructions (Fischer and van der Leek 1983: 349–352; also EXPERIENCER-object constructions in Croft 1991: 219). This is the path followed by *ail* or *please* (e.g. *Her decision pleased me*. Fischer and van der Leek 1983: 363).

## 3. SYNTACTIC AND SEMANTIC PROPERTIES OF VERBS OF DESIRE IN PDE

Levin's (1993: 194–195) class of PDE verbs of DESIRE comprises (im)personal verbs such as *crave*, *desire*, *need* or *yearn*. According to Levin, the first argument of verbs of DESIRE, i.e. "the person that desires something," may be considered a type of EXPERIENCER, which is invariably encoded as the subject of the clause. The class may be further subdivided into *want*-verbs and *long*-verbs, depending on whether the second argument, i.e. "the thing desired" (Levin 1993: 194), is encoded by a direct object as in (4) —*want*-verbs— or by a prepositional object as in (5) —*long*-verbs (examples from Levin 1993: 194–195).

[3] ME personal pronouns retain the subjective/objective case distinction in the majority of cases (Allen 1986: 378; e.g. *ic/mē*, *wē/us*, *hē/him*, etc.), except for the neuter *(h)it/(h)it*. Notice, in addition, that case distinctions were more pervasive in ME than in PDE, since the ME second-person plural form *ge* still maintained the distinction between *ge* 'ye' (subjective) and *eow* 'you' (objective) (Allen 1986: 378, fn. 2).

The particular case of *lust* is representative of the *long*-class, as it has developed prepositional uses in PDE (e.g. *pregnant women **lusting** for pickles and ice cream*; see Section 1).

(4) Dorothy **needs** new shoes.

(5) Dana **longs** for a sunny day.

Verbs of DESIRE are two-place predicates with the semantic frame <EXPERIENCER, CAUSE>. In order to characterise the semantic properties of the participants of verbs of DESIRE, the present study makes use of Dowty's (1991: 576) concept of Proto-role, which conceives semantic roles as prototypical categories formed by clusters of semantic features. Dowty (1991: 551) also introduces the 'Argument Selection Principle', which rests on the assumption that the argument that shows the greatest number of so-called Proto-agent properties will be encoded as subject, whereas the argument with the greatest number of Proto-patient properties will be encoded as direct object. The semantic clusters of features that characterise the Proto-agent and Proto-patient roles are displayed in Table 1.[4]

| | **Proto-agent** | **Proto-patient** |
|---|---|---|
| **1.** | Volitional involvement in the event or state | Undergoes change of state |
| **2.** | Sentience (and/or perception) | Incremental THEME |
| **3.** | Causing an event or change of state in another participant | Causally affected by another participant |
| **4.** | Movement (relative to the position of another participant) | Stationary (relative to movement of another participant) |
| **5.** | Exists independently of the event named by the verb | Does not exist independently of the event, or not at all |

Table 1: Semantic features of the Proto-agent and Proto-patient roles (adapted from Dowty 1991: 572)

In the framework of Dowty (1991), it becomes apparent that the CAUSE argument of verbs of DESIRE lacks the majority of Proto-patient properties. Thus, the CAUSE does not undergo a change of state (Property 1), it is not an incremental THEME (Property 2), it is not causally affected by another participant (Property 3), it does not lack movement relative to the position of another participant (Property 4) and it does have independent existence from the event named by the verb (Property 5).

[4] An incremental THEME is "an NP that can determine the aspect of the sentence, since the parts of the event correspond to parts of the NP referent that are affected by the action; the event is 'complete' only if all parts of the NP referent are affected (or effected)" (Dowty 1991: 588; cf. Hopper and Thompson's (1980: 252–253) 'affectedness of object'). An example of incremental THEME would be, for instance, the NP in the sentence *mow the lawn*, where the telic aspect of the event of mowing can be deduced from "whether the grass on the lawn is all tall, partly short, or all short" (Dowty 1991: 267).

## 4. DATA SOURCES AND METHODOLOGY

The present study is based on data drawn from EEBOCorp 1.0, an offline version of *Early English Books Online* (EEBO), which comprises works printed between 1473 and 1700 in subject areas such as English literature, linguistics, theology or fine arts. EEBOCorp 1.0 largely reproduces the database provided by EEBO, and thus includes all texts in EEBO Phase I with no genre, wordcount balance or codification for text type or subject domain. EEBOCorp 1.0 excludes non-English as well as posthumous texts, and it also filters out translations from works by long-deceased authors, even if these are not posthumous from the point of view of the translator.[5]

The size of the corpus is extremely large (525 million words) and it is perhaps not always ideal for research on frequent items, given the high number of hits retrieved whenever homonyms are involved, as happens in the present case with the verb *lust* versus the noun *lust*. In order to work with a manageable number of hits, a random selection of texts has been made totalling c. 20 million words, and including only texts written in prose. As laid out in Table 2, the corpus is structured into four sub-corpora of comparable size, across four 50-year subperiods.

| Subperiod | No. of texts | No. of words |
|---|---|---|
| S1 (1500–1549) | 200 | 4,969,243 |
| S2 (1550–1599) | 226 | 4,997,385 |
| S3 (1600–1649) | 230 | 5,003,071 |
| S4 (1650–1700) | 235 | 4,929,578 |
| Total | 891 | 19,899,277 |

Table 2: Number of texts and wordcount per 50-year subperiod

The dataset of examples with *lust* retrieved from these 891 texts consists of 273 occurrences. The selected software tool for data search is the concordancer *AntConc* (Anthony 2019). In order to identify the array of spelling variants for this verb, I first gathered the list of possible spellings provided in the OED and then checked them against the corpus word list from EEBOCorp 1.0 generated with *AntConc*. The ensuing syntactic analysis was carried out by annotating the data for factors concerning the types of clause where *lust* occurs (e.g. 'type of complementation pattern', 'main/subordinate clause', 'type of subordinate clause') and the formal realisation of arguments (e.g. 'formal realisation of the EXPERIENCER/CAUSE', 'preposition', 'personal pronoun'). For its part,

[5] I am grateful to Peter Petré for giving me access to this corpus.

the semantic analysis considered the Proto-role properties postulated by Dowty (1991) and outlined in Section 3 above. The data were annotated for factors related mainly to the features of volition (Property 1, 'volitional/unvolitional'), the feature of sentience (Property 2, 'sentient/non-sentient'), the feature of causation (Property 3, 'causation/no causation'), the feature of movement relative to the position of another participant (Property 4, 'movement/no movement') and the feature of existence independently of the event (Property 5, 'existence/no existence').

## 5. ORIGIN AND COMPLEMENTATION PATTERNS OF *LUST*

In this section, I look at the origin and development of *lust*, from ME *lusten*. The MED first documents *lust* in c1175 (?OE), although the original text presumably dates back to the OE period (MED s.v. *lusten* v. 1. [d]). The history of *lust* prior to ME is uncertain. According to the OED, *lust* derives from the noun *lust* 'pleasure, delight', a word inherited from Germanic, and it is first attested in the early thirteenth century (OED, s.v. *lust*, v. †1. a.). The MED, however, states that *lust* originates from both the ME noun *lust* (OE *lust*) and the OE verb *lystan* 'to desire', the most frequent verb in OE impersonal patterns with genitive or prepositional complements (i.e. Allen's 'Type N', 1995: 70–71; see also Möhlig-Falke 2012: 115). The present study follows Miura (2015: 62) in treating *lust-* and *list-* forms as separate lexical items on the grounds that the OED and the MED give them separate entries.

Judging from the OED, the MED and previous studies, the impersonal use of *lust* is first documented in the twelfth century in texts which were (presumably) composed in the OE period (see e.g. MED s.v. *lusten* v. 1. [a] and [d]), and is last attested in the mid-sixteenth century (see OED s.v. *lust*, v. †2.).[6] In impersonal use, *lust* is found with a pronominal EXPERIENCER in the objective case in combination with three different types of complements representing the semantic role of CAUSE: 1) CAUSE as clausal complement (see example (2) above); 2) CAUSE as prepositional complement (e.g. a1393. *Hem* ***lusteth*** *<u>of no ladi chiere</u>* 'They do not desire the countenance of a lady');[7] and 3) CAUSE as zero complement (e.g. c1475 [1392]. *By cause of heete him* ***lustiþ*** *myche* 'Because of the heat he feels a great longing').

[6] Unless otherwise stated, examples in this section are taken from the OED and the MED entries.
[7] I thank Ayumi Miura for her helpful opinion on the interpretation of this example.

A variant of impersonal patterns with clausal complements can be found in subordinate clauses where a proposition is omitted but retrievable from the preceding context; this is signalled by the empty brackets in the following examples: *as him lusteth [ ], when him lusteth [ ]*, to be compared with analogous structures taking an explicit clausal complement (e.g. c1390 [?c1350]. *Whon þe* ***lust*** *speke with me*). Similar constructions are observed since OE times, for instance with the OE verb *lystan* (e.g. *eal þaet hine* ***lysteþ*** 'all that he likes', from Elmer 1981: 117 [my translation]) or *lician* (e.g. OE *... þe estað heom silfum swa heom betst* ***licað*** ... 'who himself lives in luxury, as pleases him best', from Möhlig-Falke 2012: 144, 205). Examples like these will be termed impersonal NO PROPs (short for 'unexpressed proposition'), after Allen (1995: 86, 257–258, 275–277), and they include subordinate clauses introduced by *as* or *what, when* and variants (e.g. *whatsoever* or *whenever*).[8]

Judging from the OED, the MED and previous studies, the personal use of *lust* first emerged in the fourteenth century (see e.g. MED s.v. *lusten* v. 2. [a]), which is about two centuries after impersonal use is first documented in the twelfth century. This suggests that in its initial stages *lust* must have been restricted to impersonal use only. In addition, this also implies that personal constructions (with EXPERIENCER subjects) emerge at a time when the loss of case distinctions was at an advanced stage, whereas the fixation of word order was at an intermediate stage (see Section 2). It is also noteworthy that the personal use of *lust* is always found in EXPERIENCER-subject constructions, but never in CAUSE-subject (see also Miura 2015: 181, especially her Table 5.33).

In personal use, three different types of complementation patterns can be found, depending on the expression of the CAUSE argument, namely: 1) patterns with clausal complements (e.g. 1586. *Insomuche as hee that neuer* ***lusted*** *to helpe others, was not nowe able to helpe himselfe*); 2) patterns with NP complements (e.g. 1653. *The Spirit and the flesh are contraries, and they* ***lust*** *contrary things*); and 3) prepositional patterns (e.g. 1563. *If we be an hungred, we* ***lust*** *for bread*).

As with impersonal constructions with NO PROPs, a variant of personal patterns with clausal complements can be found in subordinate clauses where a proposition is

[8] NO PROP constructions are considered to omit a clausal rather than an NP complement because: 1) a clausal complement can be inserted without affecting the grammaticality of the clause (cf. [1404]. *Alle his Justices and his Sergeantz and othir suche as hym* ***lust*** *name*); and 2) evidence has been found of analogous structures taking an explicit clausal complement, but not an NP complement (?*as him lusteth* NP; ?*when him lusteth* NP).

understood (e.g. 1526. *They..have done vnto him whatsoever they* ***lusted*** *[ ]*). These will be termed as personal NO PROPs, in parallel to Allen's (1995) impersonal NO PROPs outlined above. In like manner, the personal NO PROP corresponds to subordinate clauses introduced by *as* or *what, when* and variants.[9]

## 6. *LUST* IN THE EMODE PERIOD

Table 3 displays the overall frequency of *lust* in EEBOCorp 1.0 distributed by 50-year subperiod and subject domain.[10] Note that the category 'General Prose' includes text types such as manuals of style or biographies which are not clearly classifiable into any of the other domains identified.

| Subject domain | S1 (1500–1549) | S2 (1550–1599) | S3 (1600–1649) | S4 (1650–1700) | Total |
|---|---|---|---|---|---|
| **Religion** | 128 (90.1) | 46 (78) | 30 (81.1) | 32 (91.4) | 236 (86.4) |
| **General Prose** | 8 (5.6) | 10 (16.9) | 4 (10.8) | 1 (2.9) | 23 (8.4) |
| **History** | 6 (4.2) | -- | 1 (2.7) | 2 (5.7) | 9 (3.3) |
| **Philosophy** | -- | 2 (3.4) | 1 (2.7) | -- | 3 (1.1) |
| **Law** | -- | -- | 1 (2.7) | -- | 1 (0.4) |
| **Politics** | -- | 1 (1.7) | -- | -- | 1 (0.4) |
| **Biology** | -- | -- | -- | -- | -- |
| **Literature** | -- | -- | -- | -- | -- |
| **Medicine** | -- | -- | -- | -- | -- |
| **Total** | 142 (100) | 59 (100) | 37 (100) | 35 (100) | 273 (100) |

Table 3: Frequency distribution of EModE *lust* by 50-year subperiod and subject domain (raw figures and percentages)

As said earlier, the search for this verb yielded a total of 273 tokens. It can be observed in the table that the overall frequency of *lust* notably decreases in the course of EModE, a diachronic picture that reflects its status as a low-frequency verb in PDE (see OED s.v. *lust*, v. Frequency [in current use]). It can also be seen that *lust* is predominantly found in religious and biblical contexts (86.4%), and this is consistent across the four subperiods.[11] In contrast, the frequency in the other subject domains is anecdotal overall, with some domains showing no attestations, such as Biology, Literature and Medicine, whereas some other domains show subperiods with no data, like Philosophy, Law and Politics.

[9] Subordinate clauses introduced by *as* or *what, when* and variants are all equally subsumed under NO PROPs because they have in common with the latter that a proposition can be added without affecting the grammaticality of the clause (e.g. *as they* ***lusted*** *[to do]*; *whatsoever they* ***lusted*** *[to do]*).

[10] Since EEBOCorp 1.0 does not provide its data coded for text type or subject domain, the classificatory labels adopted here have been devised specifically for the purposes of this research, based on the information gleaned from the text files themselves.

[11] Note that religious discourse is generally characterised by the use of archaic language, which is highly dependent on Latin (see e.g. Görlach 1993: 164–165).

In the following sections, the impersonal and personal uses found in the corpus are discussed. Section 6.1 focuses on impersonal patterns, while Section 6.2 looks at personal ones; but, before we delve into the discussion, a note needs to be made regarding their overall distribution. In the data from the EModE period, the impersonal use of *lust* is documented in just 6.2% of total occurrences (17 tokens). Impersonal patterns are therefore the least frequent use as against personal constructions, with a ratio of 1 impersonal to 15 personal instances, and, crucially, all the impersonal instances are attested in the earliest period S1 (1500–1549). It is noteworthy that, even though personal constructions start to be recorded only from the fourteenth century (see Section 5), in the EModE period personal patterns already represent 93.8% of cases (256 tokens), which may be taken as an indication that the shift from impersonal to personal use must have taken place during the late ME period.

### *6.1. Impersonal patterns in EModE*

The data in Table 4 show that, with regard to the formal realisation of the EXPERIENCER argument, unambiguous impersonal patterns have been attested only with pronominal EXPERIENCERS, namely *me* (example (9)), *thee/the* (example (7)), *him/hym* (example (6)) and *them* (example (8)).[12] With respect to the formal realisation of the CAUSE, in the EModE data it is realised by a *to*-infinitive clause (example (6)) or a NO PROP (examples (7)–(9)); no examples have been attested with either zero complements or prepositional phrases.

[12] Instances with nominal EXPERIENCERS have been counted as personal for two main reasons: 1) nouns are uninflected for case in the data here examined; and 2) the rigidification of word order was well advanced by the EModE period (see Section 2). It thus seems reasonable to assume that uninflected nominal EXPERIENCERS in preverbal position functioned as grammatical subjects in the period of study. As for the second-person plural pronouns *ye/you*, even though these retain case distinctions for the most part of the sixteenth century (Barber 1997: 149), no instances have been found where the originally objective *you* form represents the EXPERIENCER of an impersonal pattern (i.e. *?you lusteth*).

| | Main clause | Subordinate clause | | | | | |
|---|---|---|---|---|---|---|---|
| **CAUSE** | | | **EXPERIENCER** | | | | |
| | | **Noun** | | | **Pronoun** | | |
| | | | *me* | *thee/the* | *him/hym* | *them* | Total |
| ***TO*-INF** | -- | -- | -- | -- | 1 | -- | 1 |
| **NO PROP** | -- | -- | 1 | 3 | 11 | 1 | 16 |
| **Total** | -- | -- | 1 | 3 | 12 | 1 | 17 |

Table 4: Distribution of morphosyntactic properties of impersonal patterns with EModE *lust* (raw figures)

(6) 1531. so that this full power shulde be able to do any dede that is possible to be done, or any thynge that hym **lustethe** to do. (D00000998431390000.txt)

(7) 1536. and that thou mayst be fre to vse thy wordes as the **lusteth**. (D00000998400370000.txt)

(8) 1539. What so euer them **lusteth**, that proudly and stubburnly they dare do. (D00000998548110000.txt)

(9) 1549. Is it not lawful for me to do as me **lusteth** with mine owne goodes? (D00000999002540000.txt)

Impersonal patterns appear predominantly in NO PROP constructions such as those exemplified in (7)–(9) above, with the latest occurrence being recorded in 1549, and illustrated in example (9); there is only one exception to NO PROPs, corresponding to the *to*-infinitive complement in the earliest example (6). The impersonal NO PROP construction seems to show a significant degree of fossilisation, for it is attested with the third-person masculine singular pronoun *him/hym* in 70.6% of cases; further, it is introduced predominantly by *as* or *what*, *when* and variants (respectively 5 and 11 tokens). Overall, the high incidence of these fossilised structures suggests that the degree of productivity of the impersonal construction was limited already in the early sixteenth century. It may be that these NO PROPs constitute a remnant of an impersonal pattern which was previously at work, but which in EModE remains only as a fossilised expression preceding the total obsolescence of impersonal patterns with this verb.

### *6.2. Personal patterns in EModE*

In the following paragraphs, the discussion focuses on the historical development of the personal patterns documented in EEBOCorp 1.0 (1500–1700), which vary in the number and nature of the arguments expressed. The complementation patterns attested in EModE

include: 1) patterns with clausal complements (example (10)); 2) patterns with NP complements (example (11)); patterns with zero complements (example (12)); and 4) prepositional patterns (example (13)).

(10) 1529. Likewise (saide he) muste thou also punisshe and chastise thy silfe yf so thou **luste** to serve god. (D00000998455470000.txt)

(11) 1538. he shoulde **luste** those thynges that lawes allow. (D00000998408250000.txt)

(12) 1548. Thou shalt not desire or **lust**. (D00000998449220000.txt)

(13) 1628. Not to look upon the wine when it giveth his colour in the glasse; his meaning is, we should not **lust** vehemently after it. (D00000222874350000.txt)

Table 5 shows the raw frequencies for each of the documented personal patterns, with percentages in brackets; in parallel, Figure 1 provides the relative frequencies distributed across the four 50-year subperiods under analysis.[13]

| Complementation pattern | S1 (1500–1549) | S2 (1550–1599) | S3 (1600–1649) | S4 (1650–1700) | Total |
|---|---|---|---|---|---|
| **Clausal** | 59 (47.2) | 22 (37.3) | 6 (16.2) | 5 (14.3) | 92 (35.9) |
| **Zero** | 32 (25.6) | 21 (35.6) | 16 (43.2) | 22 (62.9) | 91 (35.5) |
| **Prepositional** | 29 (23.2) | 15 (25.4) | 14 (37.8) | 7 (20) | 65 (25.4) |
| **NP** | 5 (4) | -- | -- | 1 (2.9) | 6 (2.3) |
| **Other** | -- | 1 (1.7) | 1 (2.7) | -- | 2 (0.8) |
| **Total** | 125 (100) | 59 (100) | 37 (100) | 35 (100) | 256 (100) |

Table 5: Frequency of personal patterns with EModE *lust* by 50-year subperiod (raw figures and percentages)

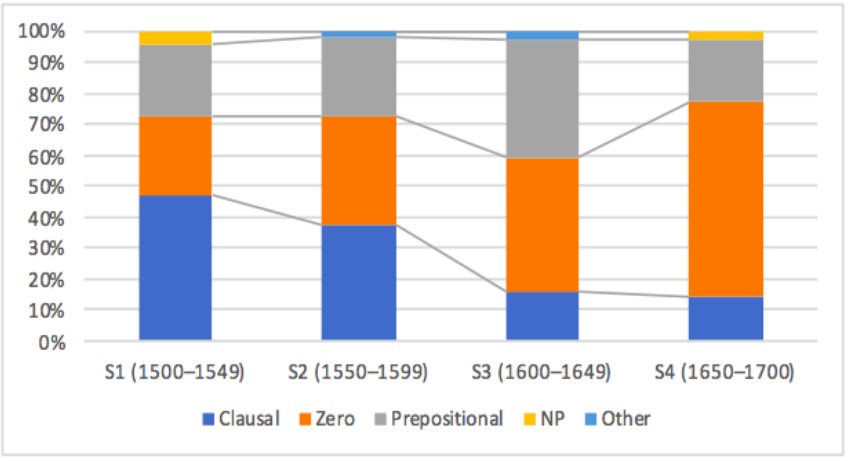


Figure 1: Diachronic distribution of personal patterns with EModE *lust* (relative frequencies)

[13] The category 'Other' includes instances of *lust* which have not been identified due to difficulties of interpretation, usually because they lack the morphosyntactic information needed for an unambiguous categorisation (e.g. 1535. *thou shall not **lust** or concupisce [Illegible_Word] desire*).

The overall relative frequencies suggest a similar distribution in clausal and zero complements, both at approximately 36%. Prepositional patterns show a lower frequency at 25.4%, while patterns with NP complements constitute the smallest proportion of occurrences at 2.3%. The diachronic evolution across subperiods reveals crucial differences with regard to the frequency of complementation patterns. Zero complements show a steady increase, rising from 25.6% of the instances in S1 to 62.9% in S4. In turn, clausal complements show a parallel decrease over time, gradually from the earliest supberiod —when they clearly dominated in frequency at 47.2%— to the final subperiod. At the turn of the sixteenth century, clausal complements undergo a marked decline from 37.3% —still competing in frequency with zero complements— to 16.2% in the early seventeenth century —i.e. less than half the frequency of the pattern with zero complements. For their part, patterns with prepositional complements remain constant except for the small increase from 25.4% in S2 to 37.8% in S3, standing below zero complements and clausal complements in S1 and S2, but only below zero complements in S3 and S4. As far as patterns with NP complements are concerned, these are modestly represented in S1 (4%), unattested in S2 and S3, and occur again anecdotally in S4 (one token, 2.9%). Overall, we may conclude that the data unveil a contrast between the sixteenth (i.e. S1 and S2) and the seventeenth centuries (i.e. S3 and S4), especially in connection with the diachronic development of clausal and prepositional complements, which show the sharpest shifts in frequency between S2 (1550–1599), S3 (1600–1649) and S4 (1650–1700).

Table 6 provides the frequencies for the formal realisation of the EXPERIENCER argument, only concerning finite clauses where a grammatical subject is overtly present. The realisation of the CAUSE argument will be dealt with in Sections 6.2.1 to 6.2.4.

| Complementation pattern | Noun | Pronoun | Total |
|---|---|---|---|
| **Clausal** | 7 (7.9) | 82 (92.1) | 89 (100) |
| **Zero** | 43 (55.8) | 34 (44.2) | 77 (100) |
| **Prepositional** | 8 (20) | 32 (80) | 40 (100) |
| **NP** | -- | 3 (100) | 3 (100) |
| **Total** | 58 (27.8) | 151 (72.2) | 209 (100) |

Table 6: Formal realisation of the EXPERIENCER argument in personal patterns with EModE *lust* (raw figures and percentages)

The data show that pronominal EXPERIENCERS are generally favoured, with 72.2% of total instances; out of 151 tokens, 128 (84.8%) represent personal pronouns distinct from relative or interrogative forms. The attested personal pronouns are the following, in order

of frequency: *they* (34 tokens, 26.6%),[14] *he* (20 tokens, 15.6%), *ye/you* (20 tokens, 15.6%), *thou* (18 tokens, 14.1%), *we* (16 tokens, 12.5%), *it* (15 tokens, 11.7%), *I* (4 tokens, 3.1%) and *she* (one token, 0.8%). The attested personal pronouns are always declinable (e.g. *they/them*, *we/us*, etc.). The second-person plural forms *ye/you* are counted as declinable because they are not used interchangeably for the most part of the sixteenth century (Barber 1997: 149). As for the neuter pronoun *it*, it is also counted as declinable because it has the old dative form *him* "as an alternative to accusative *it* all through the sixteenth century" (Barber 1997: 150).

Patterns with clausal and prepositional complements clearly prefer pronominal EXPERIENCERS, which respectively represent 92.1% and 80% of total instances, as is the case also with NP complements, which occur exclusively with pronominal EXPERIENCERS. On the other hand, zero complements are the only realisation that prefers nominal EXPERIENCERS, slightly over pronominal forms at 55.8%. This may well be due to the high incidence of the fossilised expression *the flesh lusts against/contrary to the spirit* and its variants, to be dealt with in Section 6.2.2. Figure 2 below shows the diachronic distribution of nominal and pronominal EXPERIENCERS. Notice that, even though pronominal EXPERIENCERS are by far the most common realisation overall, from a diachronic perspective nominal EXPERIENCERS undergo a considerable increase throughout the period.

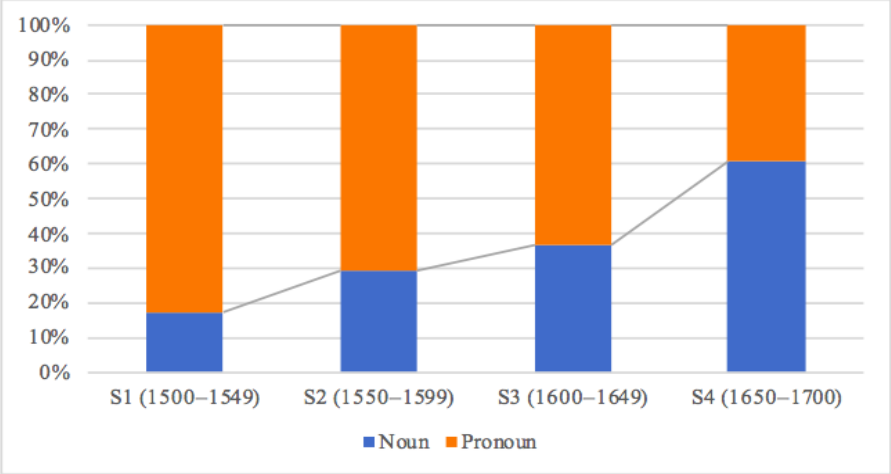


Figure 2: Diachronic distribution of nominal and pronominal EXPERIENCERS in personal patterns with EModE *lust* (relative frequencies)

[14] Note that the percentage of personal pronouns is calculated relative to the total number of personal pronouns, i.e. 128 tokens.

### 6.2.1. Patterns with clausal complements

In patterns with clausal complements, a (pro)nominal argument in the subjective case expresses the semantic role of EXPERIENCER and a clausal complement expresses the semantic role of CAUSE. According to the MED, personal patterns with clausal complements are first attested with *lust* in the fourteenth century (s.v. *lusten* v. 2. [a]). Table 7 provides the distribution for each of the realisations attested.

| CAUSE | S1 (1500–1549) | S2 (1550–1599) | S3 (1600–1649) | S4 (1650–1700) | Total |
|---|---|---|---|---|---|
| **NO PROP** | 40 (67.8) | 19 (86.4) | 2 (33.3) | 1 (20) | 62 (67.4) |
| ***TO*-INF** | 19 (32.2) | 3 (13.6) | 4 (66.7) | 4 (80) | 30 (32.6) |
| **TOTAL** | 59 (100) | 22 (100) | 6 (100) | 5 (100) | 92 (100) |

Table 7: Formal realisation of clausal complements in personal patterns with EModE *lust* (raw figures and percentages)

(14) 1528. there is no strength in their membres to doo that which their herte **lusteth** to do. (D00000998406010000.txt)

(15) 1528. What reason is it that myne enimy shulde put me in prison at his pleasure and there diet me and handyll me as he **lusteth**. (D00000998406010000.txt)

(16) 1539. Honour is offered vs, and suche honour vndoubtedly as neuer came to our nation, if we **lust** to take it. (D00000998400250000.txt)

The category of NO PROPs yields the highest number of instances (62 tokens, 67.4%; example (15)), followed by *to*-infinitive complements (30 tokens, 32.61%; examples (14) and (16)). These two variants show differences with regard to the realisation of the EXPERIENCER argument; thus, NO PROPs are found with nominal EXPERIENCERS in just one out of 62 cases, whereas the variant with a *to*-infinitive occurs with nominal EXPERIENCERS in 6 out of 30 cases; hence, the construction with a *to*-infinitive accounts for 6 out of 7 total nominal EXPERIENCERS (cf. Table 6, Section 6.2). Further, these two formal realisations show differences with regard to distribution by clause type: whereas NO PROPs are by default confined to subordinate clauses (62 tokens), *to*-infinitive complements may occasionally occur in main clauses (3 tokens). From a diachronic perspective, it is also noteworthy that the variant with NO PROPs predominates in the two subperiods of the sixteenth century, whereas the variant with *to*-infinitive does so in the two subperiods of the seventeenth century. That is, NO PROPs undergo an increase in S2 at 86.4%, but they decrease dramatically in S3 at 33.3% and in S4 at 20% (see Table 7 above). In parallel, *to*-infinitives decrease abruptly after S1 to 13.6%, but they rise up again to 66.7% in S3 and 80% in S4, though only with 4 tokens each. It can thus be

observed that a change takes place in the clausal complementation of *lust* since, as clausal complements become less frequent overall (see Table 5 and Figure 1), they also tend to disfavour an implicit realisation of the CAUSE.

Personal NO PROPs most likely represent a development of the impersonal NO PROPs discussed in Section 6.1. Personal NO PROPs share with impersonal NO PROPs that they both show a clear preference for pronominal subjects. A closer analysis of examples has revealed that in (17) below the personal NO PROP *what they lust* co-occurs with the impersonal *what him lusteth*. Notice that the personal variant takes the third-person plural pronoun *they*, whereas the impersonal variant takes the third-person masculine singular pronoun *him*, as it often happens with the impersonal NO PROPs found in the corpus. In order to assess the co-occurrence properties of personal and impersonal NO PROPs, I have examined the 14 texts where impersonal patterns have been found. It turns out that where the impersonal NO PROP takes the third-person pronoun *him/hym*, the personal variant with *he* tends to be absent. In fact, impersonal NO PROPs with *him/hym* co-occur with personal NO PROPs with *he* in only one out of the 14 texts examined. In this particular text, I found one token of impersonal NO PROP, corresponding to example (17) below, alongside 4 tokens of personal NO PROP with *he* (e.g. 1528. *and with him is lawfull what <u>he</u>* ***lusteth***). This indicates that the impersonal NO PROP with *him/hym* is retained in this context instead of the personal counterpart, which might be seen as an effect of the high degree of fossilisation and the low degree of subject variation of the impersonal patterns found in the corpus.

(17) 1528. Then love I my most enimie Now when we saye every man hath his fre will to doo what him **lusteth** I saye verely that men doo what they lust. (D00000998406010000)

From a semantic perspective, it should be remembered that in impersonal constructions the EXPERIENCER was said to denote a human being who is "unvolitionally involved in the state of affairs" (McCawley 1976: 194). Surprisingly, however, this is not the function observed in example (17) above, where the EXPERIENCER (*every man*) is explicitly said to have *his fre will* to do what he deliberately *lusteth* [i.e. chooses] to do. A similar function has been identified for personal NO PROPs in Allen (1995: 339), who points out that the personal NO PROP with *please* entails that "the Experiencer […] is in control of the action of the main clause […]." Hence, in a NO PROP construction like *I'll stay as late as I please*, the subject *I* is given the freedom to stay for as long as he or she wishes to stay, emphasising the volitional nature of the EXPERIENCER, similarly to the impersonal

construction in (17). On the face of it, it is unexpected that the impersonal and the personal variant of NO PROPs may equally attribute freedom of choice to the EXPERIENCER, when impersonal constructions have been said to denote a human being who is unintentionally involved in the event. It also seems paradoxical that impersonal constructions survive the longest in a construction type which, in the personal counterpart, contradicts their original OE function. A possible motivation for this is that the original function of impersonal constructions may have become vague in the EModE period. Going along with this, Miura (2015: 29) also observes for the ME period that "functional distinctions between the two [i.e. impersonal and personal] constructions were not always alive in late Middle English at least."

### 6.2.2. Patterns with zero complements

In patterns with zero complements, a noun or pronoun in the subjective case expresses the semantic role of EXPERIENCER, which is the only verbal argument. Since *lust* is a two-place predicate (see Section 3), it would therefore be expected to occur predominantly in clauses with two explicit arguments (cf. the OED and the MED entries). In the corpus, this is the case in the two subperiods of the sixteenth century, when clausal patterns are the preferred option; however, from the turn of the century onwards, zero complements rise in frequency and become prevalent (see Table 5 and Figure 1). Examples (18)–(20) illustrate zero complements with *lust*.

(18) 1528. The law whe~ it co~maundeth that thou shalt not **lust**. (D00000998406070000.txt)

(19) 1528. the whole nature of ma~ is damnyd in that y^ hert **lusteth** co~trary to y^ will of God. (D00000998406070000.txt)

(20) 1535. For the fleshe **lusteth** continualy agenst the sprite. (D00000998394710000.txt)

The CAUSE is left unexpressed in all of (18)–(20) above, although it may be understood as referring to the notion of sin from the religious context in which the verb is found, especially in example (18): 'thou shalt not lust [to sin/after sin]'. In (18), there are no adverbial elements in the clause and the verb denotes a '[s]ensuous appetite or desire, considered as sinful or leading to sin' (OED s.v. *lust*, n. 3.). In example (19), by contrast, an adjunct PP is introduced headed by the complex preposition *contrary to*. Likewise, a PP adjunct is present in (20), but in this last case it forms part of the fossilised expression

*the flesh lusts against/contrary to the spirit*, which expresses the meaning 'the body has carnal desires contrary to the spirit'. Alongside the variant *the spirit lusts against/contrary to the flesh*, this expression accounts for 26.4% of the total of zero complements (24 tokens).

PP adjuncts headed by *against/contrary to* are not documented in the OED or the MED. Similar uses, however, are documented with the near-synonymous verb *covet* (e.g. c1386. *The flessh* ***coueiteth*** *agayn the spirit*, defined as 'to lust' in the OED, s.v. *covet*, v. †4. a.), with which *lust* appears in coordination on two occasions in the EModE data, as in (21) below.

(21) 1538. where as the flesh coueteth and **lusteth** agaynst the spiryte.
(D00000239970610000.txt)

The high overall frequency of patterns with zero complements is partly due to the fact that they are notably represented by this formulaic expression, which is typical of the religious domain where *lust* is most common (see Table 3, especially the 'Total' for Religion). We can infer then that patterns with zero complements might not be highly productive, which is further supported if we take into account that the range of lexical nouns in subject function is considerably narrow. There are 7 different types of noun which add up to a total of 43 tokens, 30 of which correspond to *flesh* and 8 to *spirit*; the type/token ratio, which is 0.16, is therefore substantially low (cf. Bybee 2010: 94, 195). It thus appears that, although zero complements are found with high frequency, the likelihood that the pattern was used at the time to create novel utterances is limited.

### 6.2.3. Prepositional patterns

In prepositional patterns, a nominal or pronominal argument in the subjective case expresses the semantic role of EXPERIENCER, while a prepositional complement expresses the semantic role of CAUSE. According to the OED, this pattern is found with *lust* from the sixteenth century onwards, but in the EModE data prepositional patterns already represent 23.2% of uses in S1 (see Table 5 and Figure 1).

In EModE, the CAUSE argument is realised by a noun in 68.2% of cases (45 tokens) and by a pronoun in 31.8% of cases (21 tokens). The prepositional phrase expressing the CAUSE has been found to be headed by the prepositions *after* (59 tokens; example (22)), *for* (5 tokens; example (23)) and *unto* (one token; example (24)). Figure 3 shows their

relative frequency across subperiods. The preposition *after* is the most frequent collocation, a trend that remains stable throughout the EModE period. The preposition *for* is attested 3 times in S1, twice in S2 and then disappears. *Unto* is attested only once in S4.

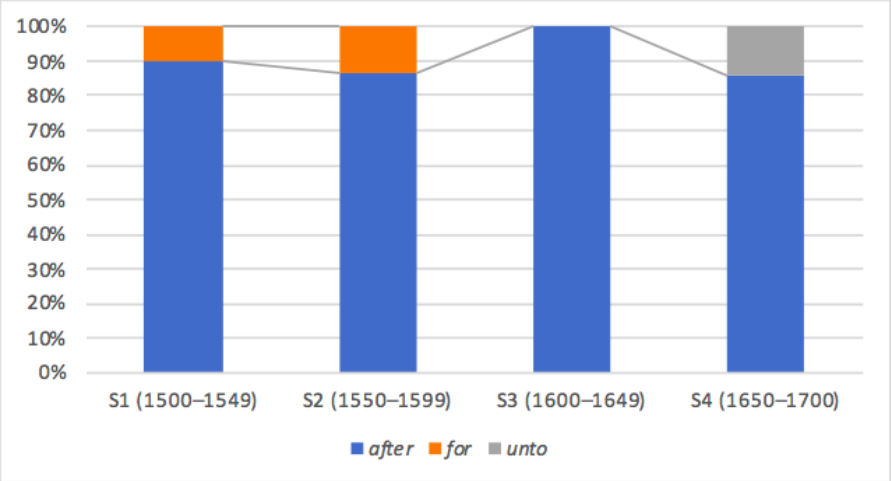


Figure 3: Diachronic distribution of prepositions governed by EModE *lust* (relative frequencies)

(22) 1543. yf my harte hathe **lusted** after my neyghbours wyfe. (D00000998461650000.txt)

(23) 1548. For we begynne to couet and **lust** for pleasant thynges, lo~g before we know whether God wyll gyue them vnto vs, or no. (D00000998449220000.txt)

(24) 1675. This the Apostle intends by its being present with us; it is present with me, that is, alwayes, and for its own end, which is to **lust** unto sin. (D00000093786480000.txt)

### 6.2.4. Patterns with NP complements

In patterns with NP complements, a nominal or pronominal argument in the subjective case expresses the semantic role of EXPERIENCER, and an NP complement expresses the semantic role of CAUSE. According to the OED, this pattern is found with *lust* only in the seventeenth century (see OED s.v. *lust*, v. †3. †d.).

In the EModE data, CAUSES are nominal in all cases (6 tokens; examples (25)–(27)).

(25) 1536. he forbade to **lust** and couet another mannes wyfe in thy harte. (D00000998400370000.txt)

(26) 1546. for these were lyght enoughe to beare, lyghter and easyer then to not **luste** or desyre any thynge agaynst goddes wyll. (D00000998386180000.txt)

(27) 1548. For who soeuer **lusteth** or desyreth in herte any thyng whiche is his neyghbours, is condemned by the law. (D00000998447910000.txt)

In example (25) the verb *lust* is coordinated with the near-synonymous verb *covet* 'to desire'. In patterns with NP complements the verb occurs in coordination with another verb in 3 out of 6 instances, including the verbs *covet* (one token; example (25)) and *desire* (2 tokens; examples (26) and (27)), both of which are amply recorded with NP complements in EModE (see OED s.v. *covet*, v. 1. a. and *desire*, v. 1. a.). *Lust* has also been found in combination with a PP adjunct headed by *against* in (26) (2 tokens), which is analogous to the PP adjuncts frequently found in patterns with zero complements.

NP complements with *lust* are very infrequent overall and, judging from the historical evidence available, they developed during the EModE period probably as an alternative to patterns with a prepositional complement for the (pro)nominal expression of the CAUSE. NP complements, however, seem to have been rapidly dismissed, probably due to the fact that CAUSES do not show any of the Proto-patient properties postulated by Dowty as contributing to the syntactic function of object (see Table 1). To recall from Section 3, CAUSES do not undergo a change of state (Property 1), they are not an incremental THEME (Property 2), they are not causally affected by another participant (Property 3), they do not lack movement relative to the position of another participant (Property 4) and they do have independent existence from the event named by the verb (Property 5). Hence, the fact that CAUSES show a low degree of affectedness (Hopper and Thompson 1980: 262) might act as a factor which makes them more eligible as prepositional than as NP complements (see Dowty 1991: 578).

## 7. SUMMARY AND CONCLUSIONS

The present case study has presented an analysis of the diachronic development of *lust* in the EModE period (1500–1700). The data have been analysed paying attention to syntactic and semantic factors. From a syntactic perspective, patterns have been characterised in terms of the formal realisation of arguments. From a semantic perspective, the properties of participants have been assessed in the light of Dowty's (1991) account of semantic roles.

With regard to the diachronic development of impersonal patterns, these have been recorded only in the first half of the sixteenth century. Considering that these have been

said to decrease in frequency between 1400 and 1500 (van der Gaaf 1904: 142; Allen 1995: 441–442), with marginal instances being found until about 1600 (Möhlig-Falke 2012: 14–15), the findings in this study are in broad agreement with the general account provided in the literature.

In connection with the occurrence of impersonal patterns after 1500, Lightfoot (1979: 229) points out that "it is more accurate to date the final obsolescence [of impersonal constructions] from the mid-sixteenth century." This claim, however, is not supported by the particular case of *lust*, since the occurrence of this pattern into EModE is largely due to its persistence in fossilised structures like *as him lusteth* and *when him lusteth*, which do not constitute instances of real productivity of the construction. The evidence, rather, is in keeping with Traugott's (1972: 130–131) observation that sixteenth-century examples represent either "conscious archaisms" or idiomatic expressions (see also Allen 1995: 279–283; Möhlig-Falke 2012: 14–15).

As to the formal realisation of arguments, the fact that pronominal EXPERIENCERS are generally favoured with this verb may follow from the fact that they are typically human beings, and "human beings are more likely to be referred to by pronouns than are things" (Allen 1995: 333). This applies to all the personal pronouns listed in Section 6.2 except the neuter *it*, which often (but not necessarily) has non-human reference. The overall high frequency of pronominal EXPERIENCERS may also be related to the general concern in religious discourse with the individual's thoughts and actions, which leads to human beings often becoming the topic of discourse. Worthy of mention is also McCawley's (1976: 198) observation that verbs that denote emotions are more likely to take pronominal EXPERIENCERS insofar as they denote "the 1st person's inherently subjective experience." Notice that this claim cannot be upheld in the present study in view of the fact that first-person pronouns are among the least frequent variants in the corpus (see Section 6.2 on the range of personal pronouns attested in personal use).

The examination of whether the arguments of *lust* are nouns or pronouns also allows us to draw some hypotheses with respect to the factors which have been claimed to affect the loss of impersonal patterns. As explained in Section 2, the reanalysis hypothesis formulated by Jespersen (1961) rests on the assumption that the SVO personal use developed from OVS sentences resembling OE *þam cynge licodon peran*, where there are two nominal NPs representing the roles of EXPERIENCER and CAUSE which eventually became morphologically ambiguous due to the loss of case inflections. As an objection

to this claim, Allen (1986) points out that the reanalysis cannot have started from this sentence type if we take into account that clauses with two nominal NPs were highly infrequent in ME data for the impersonal verb *like* (Allen 1986: 378).

In line with Allen's argument, if ambiguous case marking had been the reason for the interpretation of EXPERIENCERS as subjects in the case of *lust*, we might as well expect a large proportion of examples to have two nominal NPs at the start of the EModE period, which is about two centuries after the shift to personal use is supposed to have started in the fourteenth century (see Section 5). However, the data examined in this study contain mostly pronouns for the expression of the EXPERIENCER argument, especially in S1 (see Figure 2), which are always declinable. In addition, we have seen that NO PROPs are the construction type where impersonal constructions survive the longest with this verb; it may also be remembered that NO PROPs have the EXPERIENCER realised by a pronoun in the great majority of cases in both impersonal and personal use. It thus seems doubtful that the reinterpretation of EXPERIENCERS as subjects in NO PROPs was triggered by the ambiguity caused by a lack of case distinctions.

However, it needs to be pointed out that formal ambiguity does not seem to be a possibility in EModE, since the fixation of word order was already well advanced at the time (see Section 2). Nonetheless, the EModE data show that the syntactic scenario which may have led to ambiguity in the preceding centuries is not frequent either. In view of this, it might be fruitful to carry out a study of (late) ME data in future work in order to ascertain whether morphological ambiguity did in fact arise as the verb shifted to personal use, and while word order was not as yet rigidified.

All in all, the evidence gathered in this study shows that the EModE period witnesses crucial changes in both the meaning and the argument structure of *lust*. In the sixteenth and seventeenth centuries not only does the verb undergo a process of semantic specialisation (see Section 1), but syntactically it also becomes less frequent in patterns with clausal complements. At the same time, zero complements become surprisingly common, which may be accounted for by the frequency of idiomatised expressions typical of the religious domain where *lust* is most common. On the other hand, NP complements are only rarely found, leaving room for prepositional complements to become the preferred alternative for the expression of (pro)nominal CAUSES, in accordance with the PDE use of this verb. Lastly, it is also interesting to note that the development of prepositional patterns ties in with the development of other (im)personal

verbs of DESIRE, such as *hanker (after)*, *long (for)*, *thirst (after)* or *yearn (for)*, which similarly joined the prepositional class of verbs in PDE (Levin 1993: 194–195).

## REFERENCES


Allen, Cynthia L. 1986. Reconsidering the history of *like*. *Journal of Linguistics* 22/2: 375–409.

Allen, Cynthia L. 1995. *Case Marking and Reanalysis: Grammatical Relations from Old to Early Modern English*. Oxford: Clarendon Press.

Anthony, Laurence. 2019. *Antconc* (version 3.5.8). Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/antconc/

Barber, Charles. 1997. *Early Modern English* (second edition). Edinburgh: Edinburgh University Press.

Bybee, Joan. 2010. *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Croft, William. 1991. *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information*. Chicago: University of Chicago Press.

Dowty, David R. 1991. Thematic proto-roles and argument selection. *Language* 67/3: 547–619.

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

Elmer, Willy. 1981. *Diachronic Grammar: The History of Old and Middle English Subjectless Constructions*. Tübingen: Niemeyer.

Fischer, Olga and Frederike C. van der Leek. 1983. The demise of the Old English impersonal construction. *Journal of Linguistics* 19/2: 337–368.

Fischer, Olga, Ans van Kemenade, Willem Koopmann and Wim van der Wurf. 2000. *The Syntax of Early English*. Cambridge: Cambridge University Press.

Görlach, Manfred. 1993[1991]. *Introduction to Early Modern English*. Cambridge: Cambridge University Press.

Hopper, Paul J. and Sandra A. Thompson. 1980. Transitivity in grammar and discourse. *Language* 56/2: 251–299.

Jespersen, Otto. 1961[1927]. *A Modern English Grammar on Historical Principles. Part III: Syntax*. London: George Allen and Unwin.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

*Lexico's Dictionary*. https://www.lexico.com/en

Lightfoot, David W. 1979. *Principles of Diachronic Syntax*. Cambridge: Cambridge University Press.

McCawley, Noriko A. 1976. From OE/ME 'impersonal' to 'personal' constructions: What is a 'subject-less' S? In Sanford B. Steever, Carol A. Walker and Salikoko S. Mufwene eds. *Papers from the Parasession on Diachronic Syntax*. Chicago: Chicago Linguistics Society, 192–204.

Méndez Naya, Belén and María José López Couso. 1997. What is really meant by *impersonal*? On *impersonal* and related terms. *Atlantis. Journal of the Spanish Association for Anglo-American Studies* 19: 185–192.

*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

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*. Oxford: Oxford University Press.

*Oxford English Dictionary Online*. https://www.oed.com

Traugott, Elizabeth C. 1972. *A History of English Syntax: A Transformational Approach to the History of English Sentence Structure*. New York: Holt, Rinehart and Winston.

Trousdale, Graeme. 2008. Words and constructions in grammaticalization: The end of the English impersonal construction. In Susan M. Fitzmaurice and Donka Minkova eds. *Studies in the History of the English Language IV: Empirical and Analytical Advances in the Study of English Language Change*. Berlin: Mouton de Gruyter, 301–326.

van der Gaaf, Willem. 1904. *The Transition from the Impersonal to the Personal Construction in Middle English*. Heidelberg: C. Winter.

Visser, Fredericus T. 1963. *An Historical Syntax of the English Language. Part 1, Syntactical Units with one Verb*. Leiden: E. J. Brill.

*Corresponding author*
Noelia Castro-Chao
Campus Norte · Av. de Castelao s/n
15782 Santiago de Compostela
e-mail: noelia.castro.chao@rai.usc.es

received: August 2019
accepted: October 2019

RiCL Research in Corpus Linguistics

# Review of Cantos-Gómez, Pascual and Moisés Almela Sánchez eds. 2018. *Lexical Collocation Analysis: Advances and Applications.* Heidelberg: Springer. ISBN: 978-3-319-92581-3. https://doi.org/10.1007/978-3-319-92582-0

Pedro A. Fuertes Olivera
University of Valladolid / Spain & University of Stellenbosch / South Africa

The promotional leaflet of the *Quantitative Methods in the Humanities and Social Sciences* book series indicates that this is a book series "designed to foster research-based conversation" based on applications of "computational analysis, statistical models, computer-based programs, and other quantitative methods." One of the books in this series is *Lexical Collocation Analysis: Advances and Applications*, edited by Dr Pascual Cantos-Gómez and Dr Moisés Almela-Sánchez (University of Murcia, Spain). The editors indicate in the Introduction that the book re-examines the borderline phenomenon of 'collocation', a concept that is subjected to different, sometimes conflicting, interpretation in linguistics. This book aims to favour an integration of perspectives that will provide a kind of standardisation of the concept of collocation, perhaps one of the most productive and difficult areas of research over the decades following the introduction of the concept, which is usually attributed to J. Firth. The editors are right to argue that collocational studies "have played a central role in the *lexicalist turn* of the last decades and in the reformulation of the boundaries between vocabulary and grammar" (v). They mention Sinclair's *idiom principle* and Hoey's *lexical priming* as good "epitomes of this tendency" and indicate that the results of these and other theoretical studies have had practical applications, especially in lexicography, second language teaching/learning, and computational linguistics.

The editors are also right that collocational analysis has enormously benefited from the incorporation of the new technologies into the tools of linguistic descriptions. Hence the book "lays special emphasis on the coupling of collocational research and

computational corpus tools" (vi). In other words, the common denominator of the papers presented in the book under review "is the use of computational corpora and quantitative techniques as a means to explore aspects of language patterning that overlap the boundaries between lexis and grammar" (vi). In sum, this book offers an up-to-date analysis of the concept of collocation as this is approached in current lexicogrammar analysis carried out under the theoretical framework of corpus studies.

The book is divided into 6 chapters. In "Is language a construction? A proposal for looking at collocations, valency, argument structure and other constructions," Thomas Herbst argues "in favour of not regarding collocation and valency as strictly discrete categories but rather seeing them as near neighbours in the lexis-grammar continuum" (1). Herbst claims that collocational and valency phenomena are better understood in terms of a rather modified concept of constructions. His conclusion is that Goldberg's credo "It's construction all the way down" should be modified to "It's collexemes (or items) all the way down" (18). This 'new credo' will especially benefit second language learners who will be involved in overcoming traditional grammar books and dictionaries. I am especially convinced that this is something that deserves more attention from both grammarians and lexicographers who must focus on presenting collexemes as central units of language and offer them the best possible description.

Chapter 2 "Bridging collocational and syntactic analysis" by Violeta Seretan proposes the "coupling of collocational and syntactic analyses" (23) because one type of analysis will benefit the other. This conclusion is emphasised after reviewing the literature on both types of analysis and surveying "the work devoted to exploiting collocational resources for syntactic parsing" (23). It is interesting to highlight the review Seretan devotes to the works "that take into account the advances made in one area to foster the other area and vice versa" (35). These works really show that we need a better understanding of the proposed coupling which will allow researchers to improve language understanding.

Sánchez-Berriel, Santana Suárez, Gutiérrez Rodríguez and Pérez Aguiar's "Network analysis techniques applied to dictionaries for identifying semantics in lexical Spanish collocations" (Chapter 3) offers an innovative proposal, which basically consists in using Hausmann's *collocates* and *bases* for complementing corpus data and dictionaries in the identification of collocations and their properties. Dictionary

definitions are typically used as "a source of information to support the results obtained by the automatic extraction of collocations from a text corpus" (39) but definitions do not offer information on other important aspects of collocations, e.g. they do not distinguish if the combination is a 'collocation' in terms of the English or Russian tradition neither do they differentiate between functional and lexical collocations. To improve the deficiencies observed, they have constructed a graph database of word relationships with which they have built a complex system. This database offers better results than relational databases, especially because the "design of the lexical database model has facilitated the use of network analysis tools that discriminate different categories of collocations, particularly functional and lexical collocations" (53). The chapter convincingly shows the approach adopted and can be an inspiration for researches who want to replicate it, perhaps using different dictionaries and corpora to those employed in this chapter.

In his chapter "Collocation graphs and networks: Selected applications," Vaclav Brezina "explains the potential of collocational graphs and networks both as a visualization tool and as an analytical technique" (vi). He indicates that the notion of collocation graphs and networks goes beyond the traditional representation of collocational relationships in tabular forms. Instead, the collocational graphs and networks is a technique that can be used in, say, (i) discourse analysis; (ii) language learning research; and (iii) lexicography. He provides three case studies of how this technique really works. For instance, in lexicography he refers to the *Sketch Engine* which implements word sketches, "i.e. collocations of a word of interest categorised according to their syntactic position" (74). He indicates that dictionaries, including dictionaries that use word sketches as a methodology, do not typically include words semantically related to the entry, for instance, they do not include conceptual relations. He claims that the inclusion of conceptual relations will be very positive as they can greatly help lexicographers in lexicographic descriptions of words "beyond the usual parameters observed in electronic lexicography" (74). He supports this claim by building collocation networks related with metaphors such as TIME IS MONEY, LOVE IS A JOURNEY and ARGUMENT IS WAR. The results of his analysis are displayed in several figures and allow him to conclude that corpora provide evidence of conceptual metaphors in everyday language use and that collocation networks "automatically identify the overlaps between collocates in multiple nodes (shared collocations)" (80).

In other words, the identification of these overlaps will allow lexicographers to identify word relationships which demonstrate that "relationships between words makes [sic] collocation networks an ideal lexicographic tool" (81).

Alexander Wahl and Stefan Th. Gries propose in "Multi-word expressions: A novel computational approach to their bottom-op statistical extraction" (Chapter 5) a data-driven bottom-up approach "to the identification/extraction of multi-word expressions in corpora" (85). They present a recursive algorithm to identify multi-word expressions (MWE) called MERGE (Multi-word Expressions from the Recursive Grouping of Elements), which is based on "the successive combination of bigrams to form word sequences of various lengths" (85). In the chapter they explain the use of their created algorithm on two corpora and test its performance for extracting MWE. The chapter is a good example of how to perform high quality research in this field. Firstly, they offer their own definition of MWE. Secondly, they explain how they extracted MWE from corpora. Thirdly, they offer an empirical evaluation of the algorithm. Finally, they discuss their results and conclude that MERGE exhibits strong similarities to humanlike knowledge of formulaic language. In other words, this chapter offers an interesting example of how to combine linguistic theory, corpus technology, and statistical knowhow to identify MWEs and work with them.

Peter Uhrig, Stefan Evert, and Thomas Proisl's "Collocation candidate extraction from dependency-annotated corpora: Exploring differences across parsers and dependency annotation schemes" (Chapter 6) evaluates several parsers on two corpora with twenty different association measures plus several frequency thresholds. For carrying out such an investigation they analyse six different types of collocations against the second edition of the *Oxford Collocation Dictionary for Students of English.* Their analysis shows that although the extraction of collocation candidates is subjected to different possibilities, they recommend the use of *spaCy*, "a robust parser with good results on all relations" (135). Regarding the association measures, they also conclude that log-likelihood works well and therefore they recommend it for collocation research. This recommendation, however, must be handled with care, especially in lexicography as lexicographers would benefit if they select "different association measures for the different relations" (135).

As a conclusion of this review, I can say that this collection of book chapters is well-selected, offers up-to-date research on collocational analysis, presents very good

examples of high-quality research, and is well-edited and proofread. I should highlight that the book is a must for those interested in collocational analysis, especially researchers interested in understanding the role played by "automatic linguistic annotation (part-of-speech tagging, syntactic parsing, etc.) and using semantic criteria to facilitate the identification of collocations" (Promotional leaflet). In addition, the book offers definitions of MWEs, focuses on them for "capturing the intricacies of the phenomenon of syntagmatic attraction" and considers that collocation and valency are "near neighbours in the lexis-grammar continuum." Finally, the book illustrates the use of quantitative methods for linguistic research as this is currently done by leading scholars in the field.

*Reviewed by*
Pedro A. Fuertes-Olivera
Facultad de Comercio
Plaza del Campus, 1
47011 Valladolid
e-mail: pedro@emp.uva.es