

# Research in Corpus Linguistics



# RiCL 8/1 (2020)

## Editors

Paula Rodríguez-Puente and Carlos Prado-Alonso

ISSN 2243-4712

<https://ricl.aelinco.es/>

RiCL

Research in  
Corpus Linguistics



Official journal of

**aelinco**

Asociación Española de Lingüística de Corpus

Articles	Pages
<b>Design of a corpus of stimuli for a psycholinguistic study of lexical ambiguity</b> Natalia López Cortés	1–16
<b>COWS-L2H: A corpus of Spanish learner writing</b> Aaron Yamada, Sam Davidson, Paloma Fernández-Mira, Agustina Carando, Kenji Sagae, Claudia Sánchez-Gutiérrez	17–32
<b>The Colonial Texts Corpus for the Digital Library of Old Spanish Texts</b> Sonia Kania, Francisco Gago Jover	33–48
<b>Designing and building SCoPE<sup>2</sup>: A spoken corpus of Brazilian Portuguese and L2-English</b> Giovani Santos	49–64
<b>Brazilian cultural markers in translation: A model for a corpus-based glossary</b> Rozane Rebechi, Stella Tagnin	65–85
<b>EusTimeML: A mark-up language for temporal information in Basque</b> Begoña Altuna, María Jesús Aranzabe, Arantza Díaz de Ilarraza	86–104
<b>Corpus analysis of engagement discourse strategies in academic presentations</b> Carolina Viera, Serena AP Williams	105–130
<b>The TAGFACT annotator and editor: A versatile tool</b> Ana Fernández-Montraveta, Hortènsia Curell, Glòria Vázquez, Irene Castellón	131–146
<b>The Primary Education Learners' English Corpus (PELEC): Design and compilation</b> Zeltia Blanco Suárez, Francisco Gallardo-del-Puerto, Evelyn Gandón-Chapela	147–163
<b>The Toledo Teacher Trainees corpus (TTT): Bridging the gap between students' narratives and corpus linguistics</b> Fátima Faya-Cerqueiro, Gema Alcaraz Mármol	164–177
<b>Book Reviews</b>	
<b>Review of Fanego, Teresa and Paula Rodríguez-Puente eds. 2019. <i>Corpus-based Research on Variation in English Legal Discourse</i>. Amsterdam: John Benjamins. ISBN: 978-9-027-20235-2. <a href="https://doi.org/10.1075/scl.91">https://doi.org/10.1075/scl.91</a></b> Christopher Williams	178–194
<b>Review of Doval, Irene and M. Teresa Sánchez Nieto eds. 2019. <i>Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications</i>. Amsterdam: John Benjamins. ISBN: 978-9-027-20234-5. <a href="https://doi.org/10.1075/scl.90">https://doi.org/10.1075/scl.90</a></b> Roberto A. Valdeón	195–200
<b>Review of Amador Moreno, Carolina P. 2019. <i>Orality in Written Texts: Using Historical Corpora to Investigate Irish English (1700-1900)</i>. London: Routledge. ISBN: 978-1-138-80234-6. <a href="https://doi.org/10.4324/9781315754321">https://doi.org/10.4324/9781315754321</a></b> Raymond Hickey	201–204

# Design of a corpus of stimuli for a psycholinguistic study of lexical ambiguity

Natalia López-Cortés  
University of Zaragoza / Spain

**Abstract** – Lexical ambiguity takes place when a word has more than one meaning. This phenomenon could therefore lead to multiple difficulties in the processing of information; however, speakers deal almost effortlessly with ambiguous units on a daily basis. In order to understand how ambiguous items are processed by speakers, a clear synchronic definition of homonymy and polysemy is needed. In this paper a methodology to gather subjective information about ambiguous words and the relation within their meanings is proposed. Based on this methodology, a corpus of Spanish stimuli is being developed: this corpus consists of words classified as monosemic, homonymous and polysemous via the subjective interpretation of Spanish speakers. This corpus could be used to conduct experimental tasks to determine the behaviour in on-line processing of items with more than one meaning, in order to later design appropriate methods of approaching this complex phenomenon from the point of view of Psycholinguistics.

**Keywords** – lexical ambiguity; homonymy; polysemy; psycholinguistics; stimuli; synchrony

## 1. INTRODUCTION<sup>1</sup>

The main goal of this paper is to explain the methodology used to develop a corpus of Spanish ambiguous and non-ambiguous words. This corpus is thought to be the basis for experimental approaches to the processing and storage of meanings in long-term memory: that is to say, the words in this corpus could be used as material for psycholinguistic research. This paper therefore gives an account of the study of lexical ambiguity from a psycholinguistic point of view, placing emphasis on the importance of having corpora of materials and stimuli that have been classified taking into account subjective interpretation of words and its meanings.

Firstly, the theoretical framework is presented in Section 2, focusing on the definition of lexical ambiguity, polysemy and homonymy, and on the already existing

---

<sup>1</sup> This research was funded by DGA and was supported by the Spanish AEI and Feder (EU) through grant FFI 2017-82460-P. I would also like to thank the anonymous reviewers whose comments have greatly improved this manuscript.



psycholinguistic analysis of words with more than one meaning. This will show that there is a gap to fill in the study of lexical ambiguity in Spanish, since material based on subjective classification of homonymy and polysemy is needed. Secondly, the methodology used to design the corpus is explained in Section 3. The biggest contribution of this corpus is the subjective classification of ambiguous words as homonymous or polysemous, as well as the classification of words as non-ambiguous. To gather this information, questionnaires were used. Variable control of these lexical units will also be presented in Section 3.2. Then, a description of the current corpus and its possible applications is provided, as well as a brief comparison of the corpus with the definition of homonymy from a lexicographic point of view, in order to prove the importance of subjective measures. Finally, some future lines of research that could be undertaken to study lexical ambiguity in Spanish are sketched in Section 5.

## 2. THEORETICAL FRAMEWORK

### 2.1. *Lexical ambiguity: Different definitions of a semantic phenomenon*

Lexical ambiguity is a linguistic phenomenon that has been broadly studied. It takes place when a single lexeme has two or more meanings, as it occurs in Spanish with *llama* ‘flame’ and *llama* ‘llama’ and in English with *rabbit*-ANIMAL and *rabbit*-MEAT. It is therefore opposed to the concept of monosemy, in which a lexical form is mapped only to one meaning.

From a diachronic point of view (i.e. when the etymological origin and historical evolution of the words are considered), two types of lexical ambiguity are usually established in the literature: homonymy and polysemy. Homonymy takes places when two different words happen to converge in a single linguistic form (e.g. the Latin word *flamma* and the Quechuan word *llama*, which converge in the Spanish word *llama*). A polysemous word is produced when a word extends its meaning to designate new realities or entities (e.g. *pluma* ‘feather’ and *pluma* ‘pen’ in Spanish). Gutiérrez Ordóñez (1989: 125) claims that homonymous words can be described in terms of a phonetic convergence, whilst polysemous items, as a result of a semantic diversification.

The different types of ambiguity are reflected in lexicography. The differences in the diachronic evolution of words are depicted in the dictionaries in two ways:

homonymous words are presented under different, separated lexical entries, whereas polysemous units are presented in a single entry, where their multiple meanings are listed.

Although all the diachronic data about lexical ambiguity is interesting, it is not pertinent when studying the processing and storage of lexical units from a psycholinguistic approach. The etymological origin of words does not correlate with the psychological interpretation of ambiguous words by speakers (López-Cortés 2019).<sup>2</sup> In other words, the historical and etymological evolution of a word does not have a psychological correlate: speakers do not need to know the etymological origin of words; as a matter of fact, they are normally ignorant of it.

It is therefore important to consider the psychological interpretation of ambiguous words when studying the phenomenon of lexical ambiguity from a psycholinguistic point of view. The reason for this is that when speakers process a word, the information they access is the one stored in their memory, and the nature of that information is subjective. According to this synchronic approach, homonymy takes place when a word possesses more than one meaning and those meanings are not related in any way. By contrast, polysemy occurs when a word implies more than one meaning but those meanings are in some way related to one another. Rodd *et al.* (2002) rightly suggest that homonymous words have different meanings and polysemous words, different senses.

## 2.2. *Psycholinguistic approach to ambiguous units*

The lexical ambiguity phenomenon has played a key role in psycholinguistic research over the last decades. The fact that a single lexical form can transmit a variety of meanings, related or unrelated to one another, arouses interest mainly when trying to understand how speakers process words. Several studies investigating this topic have been carried out. The most common task to study the processing of lexical units is lexical decision tasks in which the participant needs to decide if the stimulus shown on the screen is a real word of their native language or a non-word (i.e. a string of letters that do not correspond to an actual word). When conducting this type of task, some authors discovered lower reaction times when processing an ambiguous stimulus (see, among others, Millins and Button 1989; Hino and Lupker 1996; Hino *et al.* 2002; Lin and Ahrens

---

<sup>2</sup> López-Cortés's (2019) only found processing phenomena such as those presented in Section 2.2 when the stimuli were classified following subjective metrics.

2010). In the last decade, this ambiguity advantage has been revised and a different behaviour for homonymy and polysemy has been identified. It has been discovered that the polysemous items were the only ones generating lower reaction times in lexical decision tasks (Rodd *et al.* 2002; Beretta *et al.* 2005; Klepousniotou and Baum 2007). Thus, the ambiguity advantage was reformulated into the polysemy advantage and the homonymy disadvantage.<sup>3</sup>

These differences in processing are interesting, especially since they are thought to point to differences in the way words are stored in the mental lexicon. If differences between homonymy and polysemy are found in lexical decision tasks, then these lexical units are being accessed differently in the mental lexicon. Many approaches to the storage of lexical ambiguity have been suggested (see Falkum and Vicente 2015 for a review). Nonetheless, the most extended model opts for a representation in separated, autonomous entries for homonymous words and a representation in a single entry for polysemous words.<sup>4</sup>

This model is consistent with the data about the processing of ambiguous words obtained in the lexical decision tasks. The unrelated meanings of homonymy are stored in separated entries of the mental lexicon and a competition for activation between them takes place during lexical access. As a result, higher reaction times are generated, and the homonymy disadvantage is explained. In contrast, when recognising a polysemous word, a single entry is accessed and consequently there is no competition for activation. This entry is richer and more complex than the one for homonymous items, since it should contain some sort of basic meaning that could be extended to express the specific senses of the word. The issue of the representation of polysemy in the mental lexicon has been a much-disputed subject within the field of psycholinguistics and there is still

---

<sup>3</sup> It is important to note that the data that proves a differential behaviour for homonymy and polysemy is usually based on English stimuli. Although these phenomena have been replicated in other languages (see, for example, Lin and Ahrens 2010), when conducting an experiment in Spanish the results are not clear. Haro *et al.* (2017a) were not able to find a difference between homonymy and polysemy in their experimental tasks. It can be therefore claimed that the issue of the processing of ambiguous units is still controversial and needs further reflection, especially if the comparison between languages is considered. Furthermore, the effects could change not only depending on the language used but also on the type of experimental task selected (Eddington and Tokowicz 2015).

<sup>4</sup> One of the most interesting things to consider when analysing this data is that the distinction between homonymy and polysemy may not be so strict; it is more likely to be somehow more gradual and less discrete. When conducting an experimental task, it is essential to determine the classification of the items and researchers need to establish criteria to do so, but since this may not be the most ecological solution the data needs to be examined critically. Here we propose one of these criteria to classify a subjective phenomenon as lexical ambiguity in an objective way (see Section 3.1).

disagreement about how this single entry is structured. The most extended approach is the *core meaning* theory (Klepousniotou and Baum 2007), although it has been strongly challenged in recent years by some authors (Foraker and Murphy 2012).

### 2.3. *Why is a corpus needed?*

All in all, the type of ambiguity of a word (i.e. whether it is polysemous or homonymous) affects its processing and storage. However, as already mentioned in Section 2.1, the type of ambiguity can be measured from diachrony (the origins of a lexical unit) or from synchrony (the interpretation of the relation between its meanings). Both criteria are not always equivalent. For instance, the word *catarata* (which can mean either ‘waterfall’ or ‘cataract’) is polysemous in Spanish since it has got one single Latin origin (*cataracta*). However, from a synchronic point of view, its meanings are interpreted as unrelated for which it can be considered homonymous.<sup>5</sup> It therefore follows that it is essential to determine which approach is needed to study the processing and the storage of ambiguous units.

We believe that subjective information is what is relevant when studying a semantic phenomenon from a psycholinguistic point of view. As it has already been mentioned in Section 2.1, speakers are normally not aware of etymology, and therefore of diachrony. As a consequence, in order to conduct experiments, subjectively-classified stimuli are needed, since that subjective information which is stored in the lexicon is what speakers need to have access to in order to communicate. Creating a corpus of these characteristics to study the behaviour of homonymy and polysemy is the main goal of the present, ongoing research.

It is important to point out that there are already some Spanish subjective corpora published. Estévez (1991) collected 214 subjectively-classified ambiguous words, which were then classified as homonymous and polysemous following the lexicographic criteria. Domínguez *et al.* (2001) focused entirely on polysemy, proposing 100 polysemous words. Gómez-Veiga *et al.* (2010) gathered information about 113 ambiguous words and different variables, such as frequency or dominance of meanings,

---

<sup>5</sup> The methodology used to gather this subjective interpretation is explained in Section 3.1.

but there was no further classification of those items considering the relation between their meanings.

The authors of all these corpora were aware of the fact that subjective metrics are the ones to consider when studying ambiguity from a psycholinguistic point of view. For instance, Domínguez *et al.* (2001: 65) claim that, although the dictionary directly offers the number of meanings (*acepciones*), that number is not psychologically relevant. However, these authors, as well as Estévez (1991), use the dictionary to determine whether a word is homonymous or polysemous and only consider subjective interpretation with regard to the number of meanings (in other words, to determine if an item is ambiguous or not). Besides, these materials lack a set of non-ambiguous words with which the ambiguous words can be compared, as already noted by Haro *et al.* (2017b).

The most recent efforts to design an ambiguity corpus are the ones by Fraga *et al.* (2017) and Haro *et al.* (2017b). The *Spanish Ambiguous Words Database* (SAW) by Fraga *et al.* (2017) is an interesting approach to the definition of ambiguity, since it seems to prove, via a meaning retrieval task, that the information contained in the dictionaries is quite similar to the meanings that speaker have stored in their lexicon. The participants of this study had to write meanings of different ambiguous words and those meanings were then compared, through a Pearson correlation, with the information in the lexicographic entries of the most common Spanish dictionary (*Diccionario de la Lengua Española*). The originality of this work is undeniable and its implications can be widely discussed.<sup>6</sup> However, the only metric taken into account to classify words as polysemous and homonymous is, once again, the lexicographic criterion, which means that the items are classified according to their etymological origin.

Haro *et al.*'s corpus (2017b) consists of 530 words. The most interesting contribution of this work is the fact that a methodology to identify homonymy and polysemy from a subjective point of view is proposed. Haro *et al.* (2017b) present two different subjective variables: NOM (number of meanings) and ROM (relatedness of meanings). The latter variable is obtained through a Likert scale: participants were asked

---

<sup>6</sup> There are some works that seem to prove the opposed view: the dictionary approach and the subjective, synchronic classification do not correlate (Haro *et al.* 2015; López-Cortés 2019). Some information regarding this topic is presented in Section 4.1. It is important to know that the objective of the aforementioned research and the one made by Fraga *et al.* (2017) are not equivalent, which could explain the contradictory results.

whether the meanings of a word were related and they had to select a value from 1 to 9. This is an effective way to determine the type of ambiguity of an ambiguous word and, most importantly, it is based on the interpretation of speakers. The methodology and the data analysis used by these authors are different from those presented here. A combination of both approaches could be ideal to expand the corpus and to gather more experimental stimuli in Spanish.<sup>7</sup>

### 3. CORPUS DESIGN

#### 3.1. Word classification

The most important part of the corpus design was the subjective classification of words, as ambiguous-monosemic and as homonymous-polysemous. Such classification was obtained by means of questionnaires which allowed us to gather subjective and synchronic data of words, which would later on be used as stimuli for psycholinguistic experiments. It must be noted that the methodology was consistent throughout the corpus design: the same type of questionnaire was used and the data were analysed following always similar criteria.

The stimuli corpus was designed based on data obtained through 21 questionnaires, filled between 2015 and 2019 by a total number of 716 native Spanish speakers who gave their explicit consent to participate in the experimental session. Each questionnaire had an average response of 34.09 answers with a standard deviation (SD) of 16.36. The current corpus, which is still being developed, has information about 336 Spanish words.

The questionnaires were designed using *GoogleForms* and consisted of 15–20 words each. The structure of this questionnaire is displayed in Figure 1.

---

<sup>7</sup> It is also important to note that all existent corpora are the result of psychological investigations and are therefore made by researchers working on this discipline. Nonetheless, it could be useful to have a linguistic basis to adequately design or interpret data related to a semantic phenomenon such as lexical ambiguity. For this reason, a corpus like the one presented here could be a good complement to previous works.

Figure 1: Screenshot of one of the questionnaires used to develop the corpus illustrating the Spanish word *barra* meaning ‘stick’ or ‘counter’ (among other meanings)

A word is presented, followed by two questions:

- (i) *Do you believe this word has one meaning or more than one meaning?*
- (ii) *In case you answered “more than one meaning,” do you believe the most common meanings of this word are related?* The possible answers to this question were *Yes, meanings are related* and *No, the meanings are very different*.

With these questionnaires two values were obtained: whether the word is monosemic or ambiguous (question 1) and whether the meanings of the words are considered to be related (polysemy) or not (homonymy) (question 2).<sup>8</sup>

The first words selected to start the corpus were taken from Gómez-Veiga *et al.* (2010). As it has been shown, these researchers did not consider the differentiation between homonym and polysemy and thus their words needed further classification. Then, some words from Haro *et al.* (2017a) and experimental material from Cueto *et al.* (1997) were also employed. However, these corpora were used as a source for material for word selection, but those items were always classified using our own methodology. In this way, total coherence in the design of the corpus was assured. Later on, as the corpus was being designed, new words were added by different means: experimental design and unexpected interpretation made by participants. All the new words were always classified through the questionnaires.

<sup>8</sup> One of our future lines of research is to perform meaning retrieval tasks (as in Fraga *et al.* 2017) in order to collect the most frequent meanings of these ambiguous units. However, at this stage of corpus design, our main goal was to determine a methodology to express, as objectively as possible, the opposition between monosemy-ambiguity and homonymy-polysemy.

The basis of the analysis procedure was to apply the same objective criteria to all subjective data obtained in the questionnaires. For a word to be included in one category (monosemy, polysemy or homonymy) a minimum agreement of 60% in the answers of all participants had to be reached. Some examples of this classification are presented next in (1) to (4).

- (1) *avestruz* ('ostrich'): monosemic with 80% of agreement
- (2) *flamenco* ('flamingo'-'flamenco'): ambiguous with 88% of agreement
- (3) *estrella* ('star'-'famous person'): polysemous with 75% of agreement
- (4) *banco* ('bank'-'bench'): homonymous with 95.3% of agreement

What is interesting about having this percentual information is that it reflects the fact that ambiguity seems to be a scale: the relation between meanings is gradual (see fn 4). Interpreting ambiguity this way also shows how subjective interpretation of lexical units defines the semantic phenomenon: the semantic information stored depends on the individual speakers, since not all of them interpret the items in a similar way. All this information can be useful when analysing experimental data.

Establishing a minimum percentage of agreement also helps eliminate those words that cannot be classified since their values do not reach the minimum percentage established, as it occurs with (5) and (6).

- (5) *carta* ('card'-'letter'-'menu'): between monosemy (44%) and ambiguity (56%)
- (6) *grano* ('grain'-'spot'): between polysemy (54.5%) and homonymy (45.5%)

These words have not been included in the corpus yet and further research is needed to properly classify them.<sup>9</sup> However, they lend further support to the idea that ambiguity is a gradual phenomenon and that it depends on the interpretation of each speaker.

### 3.2. *Participants and procedure*

A total number of 716 Spanish native speakers took part in the filling in of the questionnaires. Since the tool *GoogleForms* allows to collect information from the questionnaires online, some participants filled in the questionnaires from their homes but most of them did it on-site. The most common profile of participant corresponded to students of the degrees of Spanish Philology and Classical Studies at the University of

---

<sup>9</sup> We believe that the most adequate approach is to study those words from a linguistic point of view: analysing the semantic features of these units may clarify why these words are hard to classify.

Zaragoza. The age range was between 18 and 25 and all of them lived in the province of Zaragoza (Aragón, Spain).

Participants were told that their answers would be used statistically and they were asked to answer according to their own interpretation as Spanish speakers. They were aware that there was no time control and that they could use as much time as they needed to fill in the questionnaire. The fact that there were no right or wrong answers was specially stressed, so that they would answer according to their own interpretation.

The questionnaire was normally presented to the participants after they had already completed another task. A soundproofed room was used and the questionnaire was filled in via a laptop with Internet connection. The duration of this session varied depending on individual speed, but it was never longer than 15 minutes.

### 3.3. *Variable control*

The effect of ambiguity has been studied for decades now and some authors have explored the possibility that there are some variables that could interact or even interfere with the processing of words with more than one meaning. Different tasks and experiments have been carried out in the last years by researchers to determine which these variables are.

There are different approaches, methodologies and points of view but the variables most commonly studied in relation with ambiguity are the ones that follow: frequency (Rubenstein *et al.* 1970; Gernsbacher 1984; Cuetos *et al.* 1997 and more recently Jager *et al.* 2016, among others), familiarity (Gernsbacher 1984), imaginability (Cuetos *et al.* 1997) and concreteness (Tokowicz and Kroll 2007; Jager and Cleland 2016). Out of all these variables, frequency is, by far, the most amply studied. Nonetheless, its effect on ambiguity is not clear: in most cases the influence of frequency interacts with the type of task or the experimental design. For this reason, the most common approach is to control this variable when conducting an experiment: that is, using items with similar frequency to make sure that the frequency is not accountable for any processing effects that arise.

The objective when designing the corpus presented in this paper was to control for all these variables, in order to have information only about the number of meanings and their relationship. In this way it can be guaranteed that if an effect is found in an experimental task it will be caused by the ambiguity values and not by other lexical or subjective variables.

The data for all these variables was extracted from different already existent corpora: relative frequency and absolute frequency<sup>10</sup> from the NIM corpus (Guasch *et al.* 2013) and familiarity, imaginability and concreteness from the EsPal corpus (Duchon *et al.* 2013). The information related to each word was included in the corpus and was later analysed in three different groups: (i) homonymy-monosemy, (ii) polysemy-monosemy, and (iii) polysemy-homonymy. This analysis was to check that there were no statistically significant differences regarding the variables that could affect the ambiguity effect.

The non-parametric Wilcoxon test was conducted to compare the variables. The level of significance ( $p$ ) was established at 0.05. However, statistically significant results were not obtained in either of the groups, as shown in Table 1, where the result of the Wilcoxon test is presented in the first column ( $V$ -stat.).<sup>11</sup> It can be therefore claimed that there are not statistically significant differences between groups regarding these variables, which means that none of them should have an effect in experimental tasks.<sup>12</sup>

	<b>Homonymy-Monosemy</b>		<b>Polysemy-Monosemy</b>		<b>Polysemy-Homonymy</b>	
	<i>V-stat.</i>	<i>p-value</i>	<i>V-stat.</i>	<i>p-value</i>	<i>V-stat.</i>	<i>p-value</i>
<i>Relative frequency</i>	1,810	0.5394	1,655	0.2085	1,824	0.5786
<i>Absolute frequency</i>	1,811	0.5421	1,655	0.2081	1,822.50	0.5743
<i>Familiarity</i>	2,171.50	0.1957	1,779	0.9804	2,218	0.1351
<i>Imaginability</i>	2,021	0.2936	1,679	0.5167	1,897	0.7624
<i>Concreteness</i>	1,558	0.1791	2,082	0.2657	1,658	0.3613

Table 1: Results of the variable control

#### 4. CORPUS DESCRIPTION

The corpus currently consists of 336 words, subjectively classified into three groups: monosemy, homonymy and polysemy. It is therefore divided in three sections: monosemic stimuli (88 words), homonymous stimuli (88 words) and polysemous stimuli (160 words).

<sup>10</sup> Relative frequency is the appearance of the word in parts per million whereas absolute frequency is the total number of appearances of the word in the corpus, as explained by Guasch *et al.* (2013).

<sup>11</sup> One anonymous reviewer suggests including a comparison between ambiguity and monosemy in Table 1. However, we do not have this data at the moment, since we are mainly interested in the processing of homonymy and polysemy. This will be, however, considered in future research.

<sup>12</sup> One of the most important steps when designing experimental tasks is controlling variables that can have an effect on the results. For this reason, these variables should always be controlled for before carrying out any tasks. The variable control presented here works for our research since all these items were used in lexical decision tasks and, as an essential part of the corpus design, we thought it interesting to show this process in the present paper. The data presented here can work as a basis, but it is highly recommendable to repeat the controlling process depending on each researcher's experimental design and objectives.

The words in each section are ordered by agreement degree (from higher to lower) and different variables for each word are then listed: frequency (obtained from Guasch *et al.* 2013), familiarity, imaginability and concreteness (obtained from Duchon *et al.* 2013). These latter variables were measured by researchers through a Likert scale, where participants had to decide how familiar, imageable or concrete a word was in a scale from 0 to 7.

The most interesting additions to this corpus are the following: firstly, the incorporation of the homonymy-polysemy classification, based on a subjective interpretation obtained through questionnaires. This data is reflected with a percentage of the agreement in the classification, which allows us to assess whether there are differential effects of processing for words that fall within the same category but have classifications that vary greatly in agreement.

Secondly, the fact that the information about reaction times is added to the corpus is also interesting. Each word is followed by the mean of the reaction times that the item produces in lexical decision tasks. This measurement is presented in milliseconds and was obtained by conducting a series of lexical decision task with the material of the corpus.<sup>13</sup>

This information (classification, agreement and reaction times) is the major contribution of this corpus. In Table 2, a summary of the all data is presented.

	<i>Homonymy</i>		<i>Polysemy</i>		<i>Monosemy</i>	
	Mean	SD	Mean	SD	Mean	SD
<i>Reaction time</i>	763.1	95.75	737.17	78.11	730.14	68.46
<i>Relative frequency</i>	57.28	99.35	85.09	154.66	63.68	104.05
<i>Logarithm</i>	1.36	0.58	1.51	0.59	1.37	0.63
<i>Absolute frequency</i>	322.43	559.28	4790	870.63	358.49	585.72
<i>Familiarity</i>	5.27	1.62	5	1.93	4.86	2.03
<i>Imaginability</i>	4.54	1.91	4.51	2.03	4.27	2
<i>Concreteness</i>	4.08	1.42	4.27	1.67	4.29	1.91

Table 2: Summary of the data presented in the corpus, with the mean of the values and their standard deviation (SD)

With all the data gathered, we believe it is relevant to point out once again how the information compiled in the dictionary is not the same as the one in the long-term memory of native speakers, at least regarding the classification of ambiguous words, such as

<sup>13</sup> This data consists of only a small sample, but it is a preliminary approach to investigating the way in which these units are processed. Other researchers, such as González-Nosti *et al.* (2014), had already collected this measure as valuable information in previous corpora.

homonymy or polysemy (as already shown by Haro *et al.* 2015 and López-Cortés 2019; but somehow contrary to Fraga *et al.* 2017).<sup>14</sup>

Homonymy has been considered to be a far less frequent phenomenon than polysemy, since it is hard for two non-related words to converge in form. However, our psychological data reveals that homonymy seems to be more common than expected when it is measured from a diachronic point of view.

In (7), the homonymous items from our corpus are presented. These words are considered by our participants to have multiple non-related meanings. The units that are also classified as homonymous in the *Diccionario de la Lengua Española* are in bold.

- (7) apéndice, artículo, banco, **banda**, **bolsa**, **borde**, **bota**, **bote**, **cabo**, **cala**, cámara, campaña, **canto**, caña, **cardenal**, carrera, caso, catarata, **celo**, chisme, **chorizo**, chuleta, clase, **coco**, **cola**, **cólera**, **colonia**, **coma**, compañía, concierto, cuadrado, **cubo**, cura, **duelo**, esposa, estación, estado, ficha, flamenco, general, genio, **golfo**, **grado**, grano, gravedad, guion, hábito, **heroína**, **jota**, ladrón, lata, **lima**, línea, **lira**, lista, **mango**, marca, marea, **media**, medio, mina, monitor, mono, **muelle**, muñeca, nota, obra, palma, papel, parábola, parte, partida, partido, pasta, pendiente, **pez**, piña, pluma, **pompa**, puesto, pupila, rana, segundo, servicio, taco, **tapa**, **tela**, tienda.

As can be seen, there are more subjectively-classified items as not having related meanings (i.e. homonymous) than would be expected if the dictionary approach was taken exclusively: out of 88 homonymous words from the present corpus only 31 are also homonymous based on their etymology, while the other 57 are reflected in the dictionary as polysemous. In other words, 57 words that are etymologically polysemous have been reinterpreted as homonymous through our questionnaires. That means that, when taking a subjective approach, homonymy increases and speakers tend to interpret words as having unrelated meanings more frequently than expected.

The data show that a corpus design that takes into account the psychological, subjective differences between types of ambiguity is indeed a useful tool.

---

<sup>14</sup> It is important to reiterate that the objective of Fraga *et al.*'s work (2017) was to check whether the meanings speakers retrieve from memory are the same as the ones that dictionaries reflect. Their results show that there is a positive correlation between these two measures. However, the differentiation between homonymy and polysemy was not taken into account, at least not in terms of relation between meanings, since these authors considered only the number of meanings and the semantic information that each lexical form gathers.

## 5. APPLICATIONS AND FUTURE LINES OF RESEARCH

This corpus is a starting point to investigate lexical ambiguity in Spanish from a psycholinguistic point of view. Once finished, it can be used to develop experimental tasks in Spanish, as it is a source for material that has been carefully controlled. Moreover, since it has been proved that ambiguity is not a homogenous phenomenon (Klepousniotou and Baum 2007), the classification of homonymy and polysemy based on subjective interpretation can be key to a robust experimental design. We believe that the most important contribution of our research is the reflection of how the different types of ambiguity could be approached through an objective measurement of subjective interpretation which allows us to obtain a scale of values.

Having a corpus based on Spanish stimuli can be key to establish whether the processing phenomena found in English (the advantage of polysemy and the disadvantage of homonymy) are also produced in other languages.

This corpus can also be the basis of a linguistic study of words with more than one meaning. One of our lines of research is to determine the nature of the relation between meanings by studying the features that characterise polysemy and homonymy. To do so, a meaning retrieval task should be carried out (see fn 10).

Further work needs to be done to expand the corpus: more ambiguous and monosemic nouns should be subjectively classified in order to design new experimental tasks that allow us to understand the processing of different meanings. It would also be interesting to start gathering new categories such as verbs or adjectives, or even items which show ambiguity within their category (as it occurs with the Spanish word *pobre* which can either be interpreted as a noun ‘a poor man’ or an adjective ‘poor’).

We believe that designing experimental material based on subjective approaches, that is, taking into account the interpretation of speakers, is the proper way to move forward if we want to fully understand the nature of the processing mechanisms related to lexical ambiguity in particular and lexical units in general.

## REFERENCES

- Beretta, Alan, Robert Fiorentino and David Poeppel. 2005. The effects of homonymy and polysemy on lexical access: An MEG study. *Cognitive Brain Research* 24/1: 57–65.
- Cuetos, Fernando, Alberto Domínguez and Manuel de Vega. 1997. El efecto polisemia: Ahora lo ves otra vez. *Cognitiva* 9/2: 175–194.
- Domínguez, Alberto, Fernando Cuetos and Manuel de Vega. 2001. 100 palabras polisémicas con sus acepciones. *Revista Electrónica de Metodología Aplicada* 6/2: 63–84.
- Duchon, Andrew, Manuel Perea, Nuria Sebastián-Gallés, Antonia Martí and Manuel Carreiras. 2013. EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods* 45/4: 1246–1258.
- Eddington, Chelsea and Natasha Tokowicz. 2015. How meaning similarity influences ambiguous word processing: The current state of the literature. *Psychonomic Bulletin & Review* 22/1: 13–37.
- Estévez, Adelina. 1991. Estudio normativo sobre ambigüedad en castellano. *Cognitiva* 3/2: 237–271.
- Falkum, Ingrid Lossius and Agustín Vicente. 2015. Polysemy: Current perspectives and approaches. *Lingua* 157: 1–16.
- Foraker, Stephani and Gregory Murphy. 2012. Polysemy in sentence comprehension: Effects of meaning dominance. *Journal of Memory and Language* 67/4: 407–425.
- Fraga, Isabel, Isabel Padrón, Manuel Perea and Montserrat Comesaña. 2017. I saw this somewhere else. The Spanish Ambiguous Words (SAW) database. *Lingua* 185: 1–10.
- Gernsbacher, Morton Ann. 1984. Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General* 113/2: 256–281.
- Gómez-Veiga, Isabel, Nuria Carriedo, Mercedes Rucián and Óscar J. Vila. 2010. Estudio normativo de ambigüedad léxica en castellano en niños y en adultos. *Psicológica* 31: 25–47.
- González-Nosti, María, Analía Barbón, Javier Rodríguez-Ferreiro and Fernando Cuetos. 2014. Effects of the psycholinguistic variables on the lexical decision task in Spanish: A study with 2765 words. *Behaviour Research Methods* 46/2: 517–525.
- Guasch, Marc, Roger Boada, Pilar Ferré and Rosa Sánchez-Casas. 2013. NIM: A Web-based Swiss Army knife to select stimuli for psycholinguistic studies. *Behavior Research Methods* 44/3: 756–771.
- Gutiérrez Ordóñez, Salvador. 1989. *Introducción a la Semántica Funcional*. Madrid: Síntesis.
- Haro, Juan, Pilar Ferré, Roger Boada and Josep Demestre. 2015. Ambiguity advantage depends on how ambiguous words are categorized. Poster presented at the *XII International Symposium of Psycholinguistics*, University of València, 3rd July, 2015.
- Haro, Juan, Josep Demestre, Roger Boada and Pilar Ferré. 2017a. ERP and behavioral effects of semantic ambiguity in a lexical decision task. *Journal of Neurolinguistics* 44: 190–202.
- Haro, Juan, Pilar Ferré, Roger Boada and Josep Demestre. 2017b. Semantic ambiguity norms for 530 Spanish words. *Applied Psycholinguistics* 38/2: 457–475.

- Hino, Yasushi and Stephen Lupker. 1996. Effects of polysemy in lexical decision and naming: An alternative to lexical access accounts. *Journal of Experimental Psychology: Human Perception and Performance* 22/6: 1331–1356.
- Hino, Yasushi, Stephen Lupker and Penny Pexman. 2002. Ambiguity and synonymy effects in lexical decision, naming and semantic categorization tasks: Interactions between orthography, phonology and semantics. *Journal of Experimental Psychology: Learning Memory and Cognition* 28/4: 686–713.
- Jager, Bernardet and Alexandra Cleland. 2016. Polysemy advantage with abstract but not concrete words. *Journal of Psycholinguistic Research* 45/1: 143–156.
- Jager, Bernardet, Matt Green and Alexandra Cleland. 2016. Polysemy in the mental lexicon: Relatedness and frequency affect representational overlap. *Language Cognition and Neuroscience* 31/3: 425–429.
- Klepousniotou, Ekaterini and Shari R. Baum. 2007. Disambiguating the ambiguity advantage effect in word recognition: An advantage for polysemous but not homonymous words. *Journal of Neurolinguistics* 20/1: 1–24.
- Lin, Chien-Jer Charles and Kathleen Ahrens. 2010. Ambiguity advantage revisited: Two meanings are better than one when accessing Chinese nouns. *Journal of Psycholinguistics Research* 39: 1–19.
- López-Cortés, Natalia. 2019. La interpretación subjetiva de la ambigüedad léxica: Una aplicación lexicográfica. *LinRed: Lingüística en la Red* 17: 1–16.
- Millis, Michelle L. and Scoti B. Button. 1989. The effect of polysemy on lexical decision time: Now you see it, now you don't. *Memory and Cognition* 17/2: 141–147.
- Real Academia Española. 2018. *Diccionario de la Lengua Española*. <http://www.rae.es/rae.html>
- Rodd, Jennifer, Gareth Gaskell and William Marslen-Wilson. 2002. Making sense of semantic ambiguity semantic competition in lexical access. *Journal of Memory and Language* 46/2: 245–266.
- Rubenstein, Herbert, Lonnie Garfield and Jane A. Millikan. 1970. Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior* 9/5: 487–494.
- Tokowicz, Natasha and Judith F. Kroll. 2007. Number of meanings and concreteness: Consequences of ambiguity within and across language. *Language and Cognitive Processes* 22/5: 727–779.

*Corresponding author*

Natalia López-Cortés  
 Departamento de Lingüística General e Hispánica  
 Facultad de Filosofía y Letras  
 Universidad de Zaragoza  
 C/ San Juan Bosco 7  
 50009, Zaragoza (Spain)  
 e-mail: natlop@unizar.es

received: October 2019  
 accepted: February 2020

# COWS-L2H: A corpus of Spanish learner writing

Aaron Yamada<sup>a</sup> - Sam Davidson<sup>b</sup> - Paloma Fernández-Mira<sup>b</sup> -  
Agustina Carando<sup>b</sup> - Kenji Sagae<sup>b</sup> - Claudia Sánchez-Gutiérrez<sup>b</sup>  
Creighton University<sup>a</sup> / United States  
University of California, Davis<sup>b</sup> / United States

**Abstract** – This paper presents the *Corpus of Written Spanish of L2 and Heritage Speakers* (COWS-L2H), a large corpus of compositions written by North American university students learning Spanish. The goals of this work are to (1) build a large corpus of Spanish learner writing that provides samples of written data from Spanish learners in the context of a North American university, (2) to contribute corpus data collected not only from second language (L2) learners of Spanish but also from learners of Spanish as a heritage language (SHL), and (3) to develop one of the few Spanish learner corpora to provide longitudinal data.

**Keywords** – L2 Spanish; Spanish as a heritage language; Learner corpus research

## 1. INTRODUCTION

Studies in the field of second language (L2) acquisition benefit from quantities of data that are large enough to aid in the analysis of L2 learning and learner language. Since the 1980s, such quantities of data have been provided by a growing number of learner corpora, or machine-readable databases of naturally produced language spoken or written by L2 learners. These learner corpora have facilitated analyses in various areas of L2 research (see Granger *et al.* 2015 for an overview). However, while corpora of L2 English are widely available, learner corpora in other languages, such as Spanish, are much less common. Granger *et al.* (2015), for example, catalog 137 total learner corpora and note that 60% are of L2 English. This distribution of learner data is incongruent with the fact that there exists a relatively high demand for learning Spanish in North America and across the globe. In 2013, for example, 51% of students enrolled in U.S. university language courses studied Spanish (American Academy of Arts and Sciences 2016) and there are over 21 million learners of L2 Spanish across the globe (Instituto Cervantes 2019). The present paper outlines the development of the *Corpus of*

*Written Spanish of L2 and Heritage Speakers (COWS-L2H)*, a learner corpus that aims to remedy this shortcoming in available resources by providing a large sample of texts written by students enrolled in Spanish courses at the University of California at Davis, a large public North American university.

In addition to the general shortage of corpus data in L2 Spanish, there are certain gaps that exist in available learner Spanish corpora that make our endeavor necessary. For instance, there are relatively few L2 Spanish longitudinal corpora that collect data from learners at different points in their learning trajectory, in comparison to cross-sectional corpora that provide snapshots of data collected from different L2 learners at different language course levels. Without longitudinal data, researchers know relatively little about how individual learners advance in their L2 from one point in time to the next. Additionally, many current L2 Spanish corpora are relatively heterogeneous at the participant level, having sampled participants from a wide variety of learning contexts (study abroad vs. classroom, high school vs. university, etc.), which puts certain limitations on the explanatory power of the data. Finally, there are still not many available corpora that collect data from students who learn Spanish as a heritage language (that is, Spanish spoken as a minority language in a society with a different dominant language) in university courses specific for that purpose. Although research in various aspects of Spanish as a heritage language (SHL) has seen a surge in recent years (see the various chapters in Pascual y Cabo 2016), scarce works have attempted to measure the development of SHL using corpus data, which is arguably due to the unavailability of the necessary resources. Importantly, this kind of data could be used to measure the development of SHL within a classroom context, advancing what is known about the effects of institutionalized language programs for learners of a heritage language.

We aim to improve the present state of available corpus resources through the development of a new corpus of short compositions written by university students of L2 Spanish and SHL enrolled in Spanish language courses at a large public North American university. The outline of this paper is as follows: we will review presently available L2 Spanish corpora in Section 2, discuss the novel contributions of COWS-L2H in Section 3, describe its make-up and the procedure used to collect data in Section 4, present data describing our initial release in Section 5, and plot some of our future steps for this resource in Section 6.

## 2. A REVIEW OF L2 SPANISH CORPORA

Several Spanish language corpora have been designed for research purposes related to studies in sociolinguistics and historical linguistics, such as the *Corpus del Español en el Sur de Arizona* (Carvalho 2012), the *Corpus del Español* (Davies 2016), the *Corpus of Mexican Spanish in Salinas, California* (Brown 2017) and the various corpora compiled by the Real Academia Española. These corpora primarily focus on the oral and/or written production of native speakers of Spanish. There are certainly far fewer available corpora built with data produced by Spanish language learners. This is perhaps due to the wider availability of native speaker text and oral data that can be collected online or in other contexts, in comparison with the relative scarcity of learner data and limited access to L2 learners. Undoubtedly, there is a need for Spanish learner corpora in order to better understand the nature of L2 learner language, to elaborate more effective teaching practices, and ultimately contribute meaningful research to an increasingly multilingual North American society.<sup>1</sup> In this section, we provide a brief overview of some of the available L2 Spanish learner corpora and their key features.

Among available written learner Spanish corpora is the *Corpus de Aprendizajes de Español* (CAES, Rojo and Palacios-Martínez 2016), which contains 570,000 words produced in written texts by learners of all levels of Spanish within the Common European Frame of Reference (CEFR, Council of Europe 2011) except C2. Writing assignments were organized by level, such that students at different proficiency levels had different writing prompts. A variety of native languages are represented, including English, French, Arabic, Portuguese, Russian, and Mandarin. One of the largest corpora of L2 Spanish data is the *Corpus Escrito del Español como L2* (CEDEL2, Lozano 2009), a corpus directed by researchers at the Autonomous University of Madrid and the University of Granada in Spain, in collaboration with several other investigators from many other universities and secondary schools. This ongoing corpus contains written compositions in L2 Spanish from over 1,000 L1 English-speaking participants at universities and high schools around the world, compiling a corpus of currently over 800,000 words, and aiming to collect a total of one million words. Participants choose to write their compositions from a selection of twelve different topics and are also asked to complete a Spanish placement test. Designed to study the development of L2 Spanish

---

<sup>1</sup> We thank an anonymous reviewer for his/her helpful suggestions regarding this section.

morphology and syntax, CEDEL2 is effectively one of the broadest and most diverse databases of L2 Spanish available in the field.

Four oral learner corpora are currently available in L2 Spanish. The *Spanish Corpus Proficiency Level Training* (Koike and Witte 2016) was developed for language teacher training. Based on the guidelines of the American Council on the Teaching of Foreign Languages (ACTFL), it is designed to help teachers assess students' proficiency levels in Spanish. It consists of 327 videotaped oral interview sessions with 38 learners whose native language was English. It is also one of the only corpora to include learners of SHL, with data from 17 participants. The *Fono.ele Corpus* (Blanco Canales 2011) is a pronunciation-focused collection of 34,316 audio-recordings of 96 learners of a variety of native languages, at all CEFR levels in Spanish except A1 and C2. The *Spanish Learner Language Oral Corpus* (SPLLOC, Mitchell *et al.* 2008) is a corpus of oral L2 Spanish data collected from native speakers of English who completed a battery of elicitation tasks, such as picture description tasks, narratives, and oral interviews. In total, this corpus contains data collected from 60 L2 Spanish learners divided into three different levels based on proficiency and institutional enrollment. The *Corpus Oral de Español como Lengua Extranjera* (CORELE, Campillos Llanos 2014) is a corpus of oral production elicited using narrative and picture description tasks among 40 learners of L2 Spanish at CEFR levels A2 and B1. These learners were of native languages including English, French, Portuguese, and Italian, among several others.

A common limitation of all the corpora described above is that they do not feature longitudinal data. Some of the few Spanish learner corpora which do so are the *Languages and Social Networks Abroad Project* corpus (LANGSNAP, Tracy-Ventura *et al.* 2016) and the *Aprescrilov* corpus (Buyse *et al.* 2016). Designed to collect learner data in and throughout study abroad sojourns, the LANGSNAP corpus contains 300,000 words produced by 27 L1-English speaking university learners who studied abroad in Spain or Mexico. These participants produced oral and written data in a variety of elicitation tasks over a period of 20 months. *Aprescrilov*, in turn, is a large corpus of written data produced by learners of L3 Spanish whose L1 was either Dutch or French, and whose L2 was either Dutch, French, or English. These learners were enrolled in the first, second, or third year of university level Spanish and wrote more than one essay per academic quarter, which is equivalent to roughly three months.

### 3. MOTIVATION FOR THE PRESENT CORPUS

Despite the considerable utility of the above corpora, they are not without certain limitations. Principally, there is a notable lack of longitudinal L2 Spanish data, here defined as data collected from participants from at least three different points in time, following Ployhart and Vandenberg (2010). While *Aprescrilov* contains longitudinal data, it does so within a very limited timespan (one academic quarter), and collects data from L1 speakers of Dutch or French learning L3 Spanish. It is thus not of great use to those interested in L1 English-speaking learners of L2 Spanish. The LANGSNAP corpus, on the other hand, collects relatively long-term longitudinal data from L2 Spanish learners, but is limited to a small number of participants. There is clearly a need for a large corpus of longitudinally collected L2 Spanish data.

Additionally, we note that many of the above reviewed corpora collected data from relatively small quantities of participants and are thus modest in size. The corpora that are comparatively large, such as CAES and CEDEL2, are also rather heterogeneous in nature. For instance, while CEDEL2 approaches one million words, it does so in collecting data from a variety of different academic institutions (over one thousand different schools and universities), which increases the variability of these data. This is perhaps disadvantageous for researchers wishing to examine the nature of L2 Spanish within specific learning contexts, such as North American universities with large Spanish language programs. Again, we see a need for a large learner corpus that features data from a numerous but relatively homogenous group of Spanish learners, particularly for researchers interested in L2 Spanish development within a canonical university Spanish language course sequence with a uniform set of instructional syllabi and learning objectives.

Lastly, we must reiterate the fact that there are very few corpora that have collected data from SHL learners. Most research in SHL is devoted to analyzing the differences between SHL learners and native speakers of Spanish, or between SHL learners and L2 Spanish learners. Little empirical research, however, has used large quantities of data to measure SHL learners' linguistic development across the course of an academic SHL program. Large amounts of corpus data collected from SHL learners are needed to fill this gap, which is particularly relevant given that more and more institutions in the United States are designing SHL courses.

In short, while there are several learner corpora in Spanish presently available to researchers, there are also certain motivations for the construction of the present corpus. COWS-L2H thus complements the current set of Spanish learner corpora in the following three ways.

(1) COWS-L2H provides longitudinal data collected from individual learners. As described below in Section 4, participants in this corpus are asked to write a total of two compositions at two separate timepoints during the academic quarter and are allowed to participate in more than one academic quarter. Thus, this corpus includes longitudinal data collected from individuals across more than one quarter, and in several cases, more than one year.

(2) Additionally, COWS-L2H limits data collection to a single academic institution. This allows for a fine-grained analysis of the grammatical and lexical development of learners who share the same instructional context, which is that of a Spanish language program in a large public North American university. Although our corpus collects data at only one university, we know exactly which textbook our participants have used, what content is covered in their course syllabus, and what pedagogical methodology is in place in their classrooms. This allows researchers to study learners' L2 as well as the relationship between the L2 and the institutional factors that form the learning context. This is an essential point in the larger-picture notion of using corpus research to advance the effectiveness of language pedagogy.

(3) Finally, COWS-L2H is one of the few Spanish corpora to include data from learners of SHL, who are enrolled in a specific language program designed to address their unique needs.

#### 4. COWS-L2H

In this section we detail the particular institutional assets at hand that help to make our resource unique, and we outline the methodology employed to collect the writing samples that make up COWS-L2H.

#### 4.1. Institutional structure and participants

The present corpus enjoys several institutional advantages that contribute positively to its goals. First, the data are being collected at the University of California at Davis whose Spanish program offers courses in L2 Spanish at three levels: Introductory (corresponding to the first-year courses titled Spanish 1, Spanish 2, and Spanish 3), Intermediate (corresponding to the fourth and fifth-quarter courses Spanish 21 and Spanish 22), and Composition (corresponding to the sixth and seventh-quarter courses Spanish 23 and 24). The learning objectives of the Introductory and Intermediate courses are largely based on communicative competence and interaction with authentic language materials in Spanish, while the Composition courses are designed with a focus on academic writing skills in Spanish. Students can take a placement exam known as Web-based Computer Placement Exam<sup>2</sup> (WebCAPE 2.0) to be placed into these language courses. Table 1 below shows the raw WebCAPE scores necessary to be placed into the corresponding language courses.

<b>WebCAPE Score</b>	<b>Course Placement</b>
Below 260	Spanish 1
260-314	Spanish 2
315-373	Spanish 3
374-423	Spanish 21
424-464	Spanish 22
464 and above	Spanish 23

Table 1: WebCAPE 2.0 Spanish raw scores and corresponding course placement

During any given quarter, a total of roughly thirty individual sections across these course levels are offered, with a maximum of twenty-five students enrolled in each section. In general terms, in each quarter there are two to three times as many sections of Introductory Spanish offered than Intermediate or Composition sections. In all, this corpus benefits from a relatively large pool of student enrollment (roughly 750 students per academic quarter) from whom data can be collected.

Additionally, the University of California at Davis is one of few North American universities to offer a multi-level program in SHL, which consists of a three-quarter series of courses denominated Spanish 31, 32, and 33. This is significant because, as Beaudrie (2012) points out, of all U.S. universities with at least 5% Hispanic

<sup>2</sup> <https://perpetualworks.com/>

enrollment, only 38% offer SHL courses, and typically at only one level. The SHL courses at this university focus on increasing the academic proficiency in Spanish of students who learned Spanish at home and grew up with a mostly colloquial knowledge of the language. In general, these students are able to communicate effectively in an informal or familiar register, but have neither been frequently exposed to more formal registers nor to global varieties of the language. Like other SHL courses, the core emphasis is on Spanish language maintenance, the acquisition of the standard variety, and the move from receptive abilities to productive proficiency (Valdés and Parra 2018). Ultimately, the series as a whole aims at developing advanced literacy, akin to language arts courses in a monolingual context (Colombi and Harrington 2012) by building vocabulary and discursive devices associated with a diversity of dialects, registers, and genres. Generally, five sections of these SHL courses are offered each quarter, serving roughly 400 students per year.

Lastly, we aim to collect data samples continuously on a quarter-by-quarter basis, for a period of at least five years. Thus, as students continue to take courses in Spanish at this university, they can continue to contribute compositions to the corpus, providing longitudinal data that would allow researchers to measure the development of individual students' Spanish as they advance from one course to the next. Students are encouraged to take these courses and advance through the Spanish language program by the requirements of their majors, many of which require them to complete the Introductory series of a foreign language. Additionally, those students who wish to complete an undergraduate degree program in Spanish are required to fulfill the entirety of the Introductory, Intermediate, and Composition sequences, or in the case of SHL learners, the Spanish as a heritage language series. Thus, the unique advantages of compiling a corpus within this university language program are emphasized: (1) this corpus benefits from a very large participant pool; (2) this participant pool is perpetually replenished with new incoming students; (3) the language program includes a series of three consecutive SHL courses; and (4) students are encouraged to participate multiple times in the corpus project quarter after quarter, permitting the study of learner language from both a cross-sectional and a longitudinal approach. We now turn to describe the nature of the learner data collected in COWS-L2H and how these data were gathered.

#### 4.2. *Composition themes and data collection*

All students enrolled in the aforementioned Spanish courses are offered extra credit as compensation for their participation in this research project. Through the course of the academic quarter, participants are asked to write a total of two compositions in Spanish that adhere to a minimum of 250 and a maximum of 500 words. Students enrolled in the Spanish 1 course are permitted to write compositions with a minimum word count of 150 words, as many of these students are true beginners in L2 Spanish.

To date, the composition data have been collected under four different themes. For the first set of compositions, collected from 2017 to 2018, participants were asked to write about *A famous person* and *A perfect vacation*. For the following set of compositions, collected from 2018 to the present, participants wrote about the themes *A special person in your life* and *A terrible story*. These composition themes are intended to be relatively broad, to allow for a wide degree of creative liberty and open-ended interpretation on the part of the writer. For the famous person theme, for example, participants have written about famous figures of the present, of the past, and even about what it means to be a famous person. The use of such broad themes thus permits the production of a wide range of verb tenses and vocabulary. Additionally, it is important to note that we wished to choose composition themes that would be accessible to learners of all proficiency levels. In other words, we wanted to implement a broadly themed writing task that learners enrolled in any course level would be able to address. Furthermore, the rationale behind the choice of these themes, and the decision to change the themes, was to allow for certain linguistic contrasts in the data collected. For example, we changed the first theme from *A perfect vacation* to *A terrible story* in order to capture a range of linguistic structures associated with relatively positive experiences, in comparison with relatively negative experiences. Following the same rationale, the second theme was changed from *A famous person* to *A special person in your life* in order to collect a range of linguistic data related to people, one of whom was comparatively more familiar or intimate to the writer than the other.

We must recognize that a potential limitation of this open-ended composition task is that only a single written genre is represented in the corpus, which may indeed affect the findings of future analyses. However, we must also note that the advantage of utilizing a single type of written task allows for more controlled analyses of these data. A plan in place for future data collection would be to adopt a more authentic writing

task, wherein instead of writing about a special person in their lives, authors could write a letter or message directly to a special person in their lives. Such an approach would not only allow for the collection of data more reflective of real-life writing tasks, but would also capture a different range of linguistic forms associated with personal address, for example.

This research protocol was approved by the Institutional Review Board at the university where the data are collected. The large-scale collection of our corpus data is made feasible through the use of the Canvas Learning Management System, an online classroom platform that is used at this university. Students who participate in the corpus project enroll in an online Canvas site, where they consent to participate in the research before providing their written samples. They read all necessary instructions regarding the tasks and then electronically submit their typed compositions. This platform organizes their submissions into a spreadsheet database accessible to the research team. Participants are given a window of one week to redact and submit each composition, at a time and place of their choosing. During this time, participants are able to see the given theme and can take as much time as they need, within the week, to write the composition. The instructions stress that participants are to write their compositions without the aid of any other person or materials. However, there is no guarantee that participants do not resort to such aids, which we recognize as a certain drawback to the online collection of such large amounts of data: it is certainly the case that more data is often noisier than less data. We do stress to participants that the quality of their writing samples will not affect the amount of extra credit they receive, nor will their language course instructors have access to these samples. Furthermore, if we find compositions that are exact copies of previously submitted compositions, they are removed from the corpus database and their authors do not receive extra credit for that quarter. We therefore do not believe that students have any clear incentive to cheat.

A period of one month separates the submission window of the first composition from the submission window of the second composition. All participants, regardless of course enrollment, write to the same themes in any given data collection window. For example, for the first data collection point of the academic quarter all participants write about *A special person in your life* and for the second data collection point all participants write about *A terrible story*. We chose a person-based topic for the first composition theme, because this is the theme that participants address during the first

data collection point. This is important because the first data collection point takes place relatively early during the academic quarter, and as such, those at lower course levels (such as the Introductory course) generally have only learned vocabulary and grammar related to personal description and family members.

These participants must additionally complete a linguistic background questionnaire, which is hosted as an electronic form within the Canvas platform. This questionnaire is completed by participants once, at the first data collection point, for every academic quarter in which they participate. The linguistic background questionnaire collects information regarding participants' age, gender, institutional course level, instructors, native language, knowledge of other languages, and experience studying abroad in Spanish-speaking countries. It also includes a brief survey asking participants to self-rate on a scale of 1 to 5 their abilities in Spanish speaking, writing, reading comprehension, and listening comprehension. All of this information is coded into the corpus database accompanying the raw composition data.

This data collection procedure is executed each quarter. Students who have already participated in the project in previous quarters, but who wish to participate again, are able and encouraged to do so. These participants write compositions to the same themes but are asked to write entirely new compositions. In other words, a given student can write two compositions in the fall academic quarter, and then write another two compositions in the following winter academic quarter, and so on and so forth. In this manner, we are able to collect longitudinal data from the same student participants, responding to the same prompts, across multiple academic quarters.

## 5. INITIAL RELEASE: DESCRIPTIVE DATA

The data in COWS-L2H have been collected over the course of eight academic quarters from 2017 to the present date. We will continue to collect data for at least the next five years. In this section we offer basic descriptive information regarding the current status of COWS-L2H. Presently, there are 1,370 unique students who have contributed data to this corpus, including 850 native (L1) English speakers, and notably, 117 L1 Chinese speakers. Several other L1 speakers that cannot be easily clustered at the moment are also represented, such as those of Vietnamese and Tagalog. In terms of the longitudinal data we have collected, 420 participants have submitted compositions in a total of at

least two quarters (for a maximum of four writing samples from each of those students), 150 have submitted compositions in at least three quarters (for a maximum of six writing samples from each student), and 38 have submitted compositions in at least four quarters (for a maximum of eight writing samples from each student). The current attrition rate from the first data collection point to the second data collection point in an academic quarter is 11.8%. Table 2 below details the number of compositions collected according to each aggregate institutional course level at this university, the total number of words collected for each aggregate course level, and the total number of participants who submitted compositions in each aggregate course level.

Course Level	No. of compositions	No. of words	No. of participants
Introductory (Spanish 1-3)	2,058	485,435	1,130
Intermediate (Spanish 21-22)	445	120,102	244
Composition (Spanish 23-24)	536	151,197	287
Heritage (Spanish 31-33)	459	130,684	244
<b>Total</b>	<b>3,498</b>	<b>887,418</b>	<b>1,905<sup>3</sup></b>

Table 2: Descriptive summary of COWS-L2H by course level

In Table 3, we outline the number of total compositions and words written to each of the four themes: *A famous person*, *A perfect vacation*, *A special person in your life*, and *A terrible story*.

Theme	No. of compositions	No. of words
<i>A famous person</i>	892	224,328
<i>A perfect vacation</i>	806	205,720
<i>A special person in your life</i>	968	239,077
<i>A terrible story</i>	832	218,293
<b>Total</b>	<b>3,498</b>	<b>887,418</b>

Table 3: Descriptive summary of COWS-L2H by theme

COWS-L2H is freely available in TXT format to all researchers under a Creative Commons license, via a GitHub repository from which researchers can freely download our data.<sup>4</sup> An updated version of the data will be made available at the end of each academic year, once that year's data has been de-identified (that is, the names of participants and student identification numbers are not released, and the data are de-

<sup>3</sup> Note that this figure recounts students who have submitted compositions across different aggregate course levels, and thus differs from the number of unique participants who have submitted compositions to the corpus, which is 1,370.

<sup>4</sup> See <https://github.com/ucdaviscl/cowsl2h>

identified by hand in such a way that it would not be possible to link them to students' university records).

## 6. LIMITATIONS AND FUTURE STEPS

As we move forward with the construction of this corpus, one of our primary goals is to attain a greater balance among the different course levels from which we are collecting data. We recognize the challenges that exist with respect to the availability of participants in course levels that are not as numerous offered and/or populated as others and we are, therefore, considering increased recruitment efforts in these areas. It is worth noting, however, that having a larger number of students at the lower proficiency levels is important in that on average they produce fewer words per composition. Additionally, as often is the case in longitudinal data collection, one of the challenges we face is attrition, in that there is no guarantee that students who participate in the corpus once will participate again during the same quarter, or across multiple quarters. We do, however, require that student participants complete both compositions during an academic quarter to receive the extra credit compensation during that quarter.

We are currently undertaking efforts to develop and implement an error-annotation procedure for errors related to gender and number agreement and the use of the Spanish preposition *a* for direct object marking (e.g. *Respeto a los ancianos* '(s)he respects the elderly'). Our hope is that future research studies on these areas of L2 Spanish grammar could benefit from the use of error-tagged data drawn from this corpus. Additionally, we aim to design and launch search tools and an online interface to facilitate the use of the corpus.

In terms of the research which can be conducted with this corpus data, we hope to undertake preliminary analyses regarding vocabulary size and to compare these results with other large corpora. Another area of investigation that will be worth exploring is the relation between students' written production and their classroom materials. Indeed, no study to date has accumulated such a large amount of written production data in a context where these data can be matched with the syllabus and textbooks that were used at the time of writing. This will help us to better understand what the impact of classroom materials actually is on the written expression of the learners. This kind of information is relevant in the development of language teaching programs, and in

testing the effect of institutional changes on the writing samples of the learners. Similarly, in terms of SHL data, COWS-L2H will allow us to (1) examine the synchronic characteristics associated with this speech community as a unique and localized variety of Spanish (Otheguy and Stern 2011); (2) track the impact of instruction on the development of academic linguistic devices and advanced literacy among heritage speakers (Colombi 2015) as these students progress through the three-course SHL series; and (3) extract the patterns associated with each level in order to create appropriate and much-needed SHL placement tests, and inform curriculum design targeting different SHL proficiencies (Beaudrie 2012).

## 7. CONCLUSION

This paper has presented COWS-L2H, a new learner corpus whose objective is to track the development of written Spanish language skills as observed over the course of a North American university Spanish language program. COWS-L2H aims to collect large amounts of longitudinal data that are currently scarce in the field of learner corpus research. COWS-L2H also collects data from students within a single homogeneous university language program, which is significant in that it provides data collected from students following a uniform set of learning objectives and pedagogical materials. Although we recognize that building a corpus at only one institution imposes certain limitations on our data, it is our hope that the research community would use our corpus in tandem with other available learner corpora. In this sense, one of our goals is to contribute to the larger resource network of learner corpora utilized by researchers seeking to draw generalizable conclusions about L2 learning in North American university settings. Finally, COWS-L2H is among the only corpora to collect large quantities of data from learners of SHL, which will provide valuable information to investigators working to advance research with respect to the development of learners enrolled in university language courses specifically designed for heritage language learners. In total, COWS-L2H is a significant step forward in the current landscape of corpus resources available to researchers working in the fields of Spanish as a second language and Spanish as a heritage language.

## REFERENCES

- Alonso-Ramos, Margarita ed. 2016. *Spanish Learner Corpus Research: Current Trends and Future Perspectives*. Amsterdam: John Benjamins.
- American Academy of Arts and Sciences. 2016. *The State of Languages in the U.S.: A Statistical Portrait*. Cambridge, Massachusetts: American Academy of Arts and Sciences.
- Beaudrie, Sara M. 2012. Introduction: Development in Spanish heritage language placement. *Heritage Language Journal. Special Issue on Spanish Assessment* 9/1: i–xi.
- Blanco Canales, Ana. 2011. *Fono.ele*, una herramienta Web para la investigación de la competencia fónica y la formación de profesores. In Carmen Hernández González, Antonio Carrasco Santana and Eva Álvarez Ramos eds. *La Red y sus Aplicaciones en la Enseñanza-Aprendizaje del Español como Lengua Extranjera*. Servicio de Publicaciones Universidad de Valladolid, 129–140.
- Brown, Earl K. 2017. *Corpus of Mexican Spanish in Salinas, California*. <http://itcdland.csumb.edu/~eabrown> (24 November, 2019.)
- Buyse, Kris, Lydia Fernández Pereda and Katrien Verveckken. 2016. The *Aprescillov* corpus, or broadening the horizon of Spanish language learning in Flanders. In Margarita Alonso-Ramos ed., 143–168.
- Campillos Llanos, Leonardo. 2014. A Spanish learner oral corpus for computer aided error analysis. *Corpora* 9/2: 207–238.
- Carvalho, Ana M. 2012–. *Corpus del Español en el Sur de Arizona (CESA)*. University of Arizona. <https://cesa.arizona.edu/> (18 February, 2020.)
- Colombi, María Cecilia. 2015. Academic and cultural literacy for heritage speakers of Spanish. A case study of Latin@ students in California. *Linguistics and Education* 32/A: 5–15.
- Colombi, María Cecilia and Joseph Harrington. 2012. Advanced biliteracy development in Spanish. In Sara M. Beaudrie and Marta Fairclough eds. *Spanish as a Heritage Language in the United States: The State of the Field*. Georgetown University Press, 241–258.
- Council of Europe. 2011. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. <https://www.coe.int/en/web/common-european-framework-reference-languages> (24 November, 2019.)
- Davies, Mark. 2016–. *Corpus del Español: Two billion words, 21 countries*. <http://www.corpusdelespanol.org> (24 November, 2019.)
- Granger, Sylviane, Gaëtanelle Gilquin and Fanny Meunier. 2015. Introduction: Learner corpus research— past, present and future. In Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier eds. *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 1–5.
- Instituto Cervantes. 2019. *El Español: Una Lengua Viva*. Madrid: Instituto Cervantes.
- Koike, Dale and Jennifer Witte. 2016. Spanish corpus proficiency level training website and corpus: An open-source, online resource for corpus linguistics studies. In Margarita Alonso-Ramos ed., 169–196.
- Lozano, Cristóbal. 2009. CEDEL2: Corpus Escrito del Español como L2. In Carmen M. Bretones José Francisco Fernández Sánchez, José Ramón Ibáñez Ibáñez, María Elena García Sánchez, María Enriqueta Cortés de los Ríos, Sagrario Salaberri Ramiro, María Soledad Cruz Martínez, Nobel Perdú Honeyman and Blasina Cantizano Márquez eds. *Applied Linguistics Now: Understanding Language and*

- Mind/La Lingüística Aplicada Actual: Comprendiendo el Lenguaje y la Mente*. Almería: Universidad de Almería, 197–212.
- Mitchell, Rosamond, Laura Domínguez, María J. Arche, Florence Myles and Emma Marsden. 2008. SPLLOC: A new database for Spanish second language acquisition research. *EuroSLA Yearbook* 8/1: 287–304.
- Otheguy, Ricardo and Nancy Stern. 2011. On so-called Spanglish. *International Journal of Bilingualism* 15/1: 85–100.
- Pascual y Cabo, Diego ed. 2016. *Advances in Spanish as a Heritage Language*. Amsterdam: John Benjamins.
- Ployhart, Robert E. and Robert J. Vandenberg. 2010. Longitudinal research: The theory, design, and analysis of change. *Journal of Management* 36/1: 94–120.
- Rojo, Guillermo and Ignacio M. Palacios-Martínez. 2016. Learner Spanish on computer: The CAES ‘Corpus de Aprendices de Español’ project. In Margarita Alonso-Ramos ed., 55–87.
- Tracy-Ventura, Nicole, Rosamond Mitchell and Kevin McManus. 2016. The LANGSNAP longitudinal learner corpus. Design and use. In Margarita Alonso-Ramos ed., 117–142.
- Valdés, Guadalupe and María Luisa Parra. 2018. Towards the development of an analytical framework for examining goals and pedagogical approaches in teaching language to heritage speakers. In Kim Potowski ed. *The Routledge Handbook of Spanish as a Heritage Language*. London: Routledge, 301–330.

*Corresponding author*

Aaron Yamada  
 Creighton University  
 Hitchcock 110C  
 2500 California Plaza  
 Omaha, NE 68178  
 United States  
 e-mail: aaronyamada@creighton.edu

received: December 2019  
 accepted: February 2019

# The *Colonial Texts Corpus* for the *Digital Library of Old Spanish Texts*

Sonia Kania<sup>a</sup> – Francisco Gago Jover<sup>b</sup>  
University of Texas at Arlington<sup>a</sup> – College of the Holy Cross<sup>b</sup> / United States

**Abstract** – This article offers a detailed description of the *Colonial Texts Corpus*, one of eleven subcorpora of the *Digital Library of Old Spanish Texts* published by the Hispanic Seminary of Medieval Studies. Launched in 2018, the corpus allows interactive access to semi-paleographic transcriptions of texts produced in the Americas during the colonial period, a textual type that is under-represented in existing electronic corpora. The rationale of the project is provided, as well as the criteria for the selection of texts to be included and their method of preparation. Finally, the interface of the corpus is illustrated, and its functionality is exemplified.

**Keywords** – electronic corpus; *Digital Library of Old Spanish Texts*; colonial texts; Colonial Spanish

## 1. INTRODUCTION

The *Colonial Texts Corpus* is one of eleven subcorpora of the *Digital Library of Old Spanish Texts* published by the Hispanic Seminary of Medieval Studies.<sup>1</sup> This paper provides an overview of the *Corpus of Colonial Texts* project, including the rationale behind its inception, the criteria established for the selection of texts, and the methodology employed in their preparation. Likewise, a brief history of the construction of the corpus is provided, as well as an illustration of its interface and examples of its functionality. Before describing the present project, it would be beneficial to contextualize it within the framework of other digital projects undertaken by the Hispanic Seminary of Medieval Studies.

---

<sup>1</sup> See <http://www.hispanicseminary.org>



## 2. BACKGROUND OF THE *DIGITAL LIBRARY OF OLD SPANISH TEXTS*<sup>2</sup>

The *Digital Library of Old Spanish Texts* (DLOST) is an online resource prepared by the Hispanic Seminary of Medieval Studies (HSMS, or the Seminary), a non-profit publisher that grew out of the Seminario de Estudios del Español Medieval. The latter was founded at the University of Wisconsin-Madison in 1931 by Professor Antonio García Solalinde, a renowned medieval philologist and disciple of Ramón Menéndez Pidal. HSMS has been a trailblazer in the use of digital technology in the humanities. In the early 1970s, then HSMS directors, Lloyd A. Kasten and John J. Nitti, began using computers as an important tool for the compilation of dictionaries and the analysis of texts. For their *Dictionary of the Old Spanish Language* project, they eschewed the use of modern editions of medieval texts as the source material, demanding that the primary sources be as free from editorial bias as possible. They created a data bank with machine-readable transcriptions of all the texts that would eventually be incorporated into the dictionary. In 1978, the HSMS published its first texts on microfiche, in what was to become the well-known *Texts and Concordances* series.

By 1997, HSMS had begun publishing the *Texts and Concordances* on CD-ROM. Although the new physical support allowed for easier access to the transcriptions (e.g. dedicated microfiche readers were no longer needed), the texts and concordances were still non-interactive flat files, which did not allow scholars to take advantage of their full range of possibilities. In 2005, the Seminary began exploring the possibility of offering all of its textual archives in an online format. These efforts culminated in the *Digital Library of Old Spanish Texts*, launched in 2011 with the publication of the *Prose Works of Alfonso X el Sabio*. This open-access repository preserves the original structure of the HSMS texts, but allows for a truly interactive access to the semi-paleographic transcriptions, as well as to a series of indexes (alphabetical, frequency, reverse alphabetical), and concordances in KWIC format.<sup>3</sup> It is to be noted that DLOST is not a digital corpus like the *Corpus Diacrónico del Español* (CORDE), for example, but rather a digital library organized into subcorpora, grouped according to author, subject, dialect, geographic region, or literary genre. Researchers are able to perform some basic linguistic

---

<sup>2</sup> This overview is based on Gago Jover (2011, 2015). Other sources are cited where appropriate.

<sup>3</sup> A Key Word in Context (KWIC) concordance is a listing of all the words that occur in a text; each key word is shown within its immediate context, i.e. with forms both to the left and to the right of the key word, with a reference to where it appears (folio and line).

searches of the contents of the texts, within individual texts or within each subcorpus.<sup>4</sup> The principal aim of DLOST is to facilitate access to the more than 400 transcriptions published by the Seminary since 1978, with the indices and concordances being the principal means of access to the texts. By 2017, ten subcorpora had been published on DLOST, representing a total of 346 texts with nearly twenty-eight million tokens of data.<sup>5</sup>

### 3. THE *COLONIAL TEXTS CORPUS*

The *Corpus of Colonial Texts* (CCT) project represents the logical next step for the *Digital Library of Old Spanish Texts*. Given the constraints of time and resources, only Peninsular medieval and early modern texts had been converted to the online format prior to the inception of the present project. HSMS' *Colonial Spanish American Series*, which includes some nine works, had not been incorporated into the repository. With the *Colonial Texts Corpus*, we intend to greatly expand the Seminary's publications related to colonial Spanish America.<sup>6</sup> We describe in detail the parameters of the corpus below and provide a brief history of its construction.

#### 3.1. *Rationale and objectives*

The goal of our project is to produce a corpus of philologically rigorous transcriptions of Spanish colonial texts and incorporate them into the Seminary's DLOST, a publication medium that will enable open, interactive access to the texts in an online format. The overarching impetus of the project is to provide reliable primary sources to inform the history of the Spanish language during the colonial period. Despite the recent advances in the availability of electronic corpora from which to extract empirical data to perform such studies, the low number of texts from Latin America included in these corpora is

---

<sup>4</sup> A lemmatized database with advanced search capabilities, which will include all HSMS texts, is in preparation. This is the *Old Spanish Textual Archive*, or OSTA (see Gago Jover and Pueyo Mena 2018a, 2018b).

<sup>5</sup> These are, in order of publication: *Prose Works of Alfonso X el Sabio*; *Spanish Medical Texts*; *Navarro-Aragonese Texts*; *Spanish Legal Texts*; *Spanish Biblical Texts*; *Spanish Poetic Texts*; *Early Celestina Texts*; *Spanish Chronicle Texts*; *Lazarillo de Tormes (1554) Texts*; *Fuero General de Navarra Texts*. Full bibliographic information can be found in Gago Jover (2011).

<sup>6</sup> As is the prevalent practice in the United States and elsewhere, we use the term 'colonial' as a descriptor relating to the territories of Latin America that maintained political ties with Spain during the period 1492 to 1898. Our use of the term is in no way pejorative, but rather a means of encompassing the wide variety of administrative structures that existed during the time period, including viceroalties, captaincies, etc. (see Bethell 2002).

striking. For example, the Real Academia Española's CORDE, a corpus which spans the beginning period of the language until 1974, contains a textual archive in which only 6% of texts are from Latin America. The texts of the *Corpus Hispánico y Americano en la Red: Textos Antiguos* (CHARTA) network, a project aimed at publishing texts from Spain and Latin America from the twelfth to the nineteenth centuries, has 8%. While we recognize that temporal and geographic criteria limit the pool of Latin American texts, even in Davies' (2002–) *Corpus del Español* only 16% of the texts dated 1500–1900 are from Latin America.<sup>7</sup> Considering the fact that 90% of Spanish speakers reside in the Americas, the lack of representative texts needs to be addressed.

In the area of Colonial Spanish studies, we are fortunate that Spain's colonizing enterprise has left us with a plethora of primary documentation. Nevertheless, many of the seminal texts from the period only reach the public via modern editions or, as is the case of the documentary record of the U.S. Hispanic Southwest, in the form of English translations (cf. Craddock 2015). This has subjected the original texts to biases, including misreadings and mistranslations. A case in point can be drawn from one of the texts of our corpus, *Relación de la Jornada de Cíbola* by Pedro de Castañeda de Najera, which offers an eyewitness account of the Coronado expedition of 1540–1542. The *Relación* survives in a copy from 1596; the classical rendition of the text is Winship (1896). While the latter's edition and translation are of obvious historical interest, the transcription conflicts with current standard philological practice in several respects. For instance, no indications are given for folio numbers in the original manuscript, abbreviations are not adequately explained, and punctuation is not included. Most importantly, there are also numerous instances of transcription errors. In the first paragraph of the text alone, there are three mistakes: *las cosas e casos* (fol. 1r6) 'the things and cases' is transcribed as *las cosas acasos* (1896:414), *aquella* (fol. 1v16) as *aquello* (415), and *no le faltara de que dar relación* (fol. 2r10–11) 'will not be lacking [material] about which to provide an account' as the nonsensical *no le faltara de quedar relación* (415).

For historical linguists, who must find their evidence in orthographic cues, even more benign editorial interventions, such as spelling modernizations, can render the texts virtually useless for their purposes. Cortés' *Cartas de Relación* provide illustrative examples of the importance of scrupulously maintaining the orthography of the primary

---

<sup>7</sup> The data presented above are taken from Company Company 2019.

text.<sup>8</sup> One of the authoritative editions of Cortés' texts is Delgado Gómez (1993). It is based on the Vienna Codex with variants noted, except those of a phonetic nature. Delgado Gómez (1993: 100–102) loosely interprets what is considered phonetic, modernizing much of the spelling, including variations between /e/ and /i/, whereby *seguio* is represented as *siguió*, between *b* and *v* (*biven* becomes *viven*), and between *ç* and *z* (*dezir* > *decir*). Likewise, the use of *h* is regularized (*artos* becomes *hartos*), double *ss* is modernized to *s*, whereby all imperfect subjunctive verbs in *-sse*, for example, are spelled *-se*, and even *x* becomes *j* (*dixeron* > *dijeron*). These changes obscure data related to some of the most important phonological developments of the language during the fifteenth and sixteenth centuries, including variation between atonic vowels, the merger of /b/ and /β/, the devoicing of the sibilants, the loss of /h/ in words that descended from Latin F-, and the retraction of the articulation of Old Spanish /ʃ/ to Modern Spanish /x/ (see Lapesa 1981; Penny 2002; Torrens Álvarez 2018). For this reason, paleographic editions, which faithfully represent the language of the originals, are more reliable.<sup>9</sup>

Equally important is the issue of accessibility—Old Spanish texts are usually preserved in libraries and archives that require special access. Even when open access to texts is provided through digital means, non-specialists are not often equipped to decipher the handwriting or typescript of the text. There is thus a critical need for faithfully edited primary sources of colonial Spanish America that can be accessed by a variety of users. In the absence of such documentary sources, we will be unable to further our knowledge of the language of the period, of its concomitant cultural manifestations, and of the history it tells.

### 3.2. *Scope: Temporal, geographic, and typological*

Texts to be included in the *Colonial Texts Corpus* will be those written in any area of the Americas during the colonial period, 1492 to Independence. Given the varied chronology of the independence movements by country, the end date will depend on the area involved, for example, 1821 for Mexico but 1898 for Cuba. Texts with an original

---

<sup>8</sup> Cortés is said to have written five *cartas de relación*, or official reports that he sent to Charles V regarding the conquest of Mexico. The first *carta* was either lost or never existed; in editions of the *Cartas de Relación*, the *Carta de Veracruz*, written by members of the town council in 1519, takes its place. The *cartas* survive in the Vienna Codex, which includes all five letters, and the Madrid codex, which includes the four *relaciones*. See Delgado Gómez (1993).

<sup>9</sup> For other examples of why we need reliable editions of colonial texts, see Craddock and Polt (2008).

production date (OPDT) and a specific production date (SPDT) that both fall within the colonial period are preferred.<sup>10</sup> Until the arrival of the printing press in Mexico in 1539 and its subsequent spread to other areas of the Americas, many early colonial texts were printed in Spain. Therefore, place of composition will be loosely construed as ‘American’ for texts that are closely related to colonial Latin America but which may have been copied or published elsewhere. This is especially relevant for texts from the sixteenth century. For example, Cortés’ *Cartas de Relación* were written in Mexico. Although the originals are lost, the texts are extant in manuscript copies (see Section 3.1). Three survive in early imprints published in Spain.<sup>11</sup> Likewise, the *Relación de la Jornada de Cíbola* was composed in San Miguel de Culiacán, Mexico, but survives in a copy produced in Seville in 1596.

Texts to be included in the corpus will be of a wide variety, both verse and prose. Although we recognize the value of archival materials for studying the historical development of the language, brief notarial documents will not form part of the corpus.<sup>12</sup> Our focus is on texts of a more extensive narrative nature, which will serve as source material not only for DLOST, but also for OSTA. The following serve as examples of the ideal types of texts to be included in the corpus: chronicles, *memoriales*, *relaciones*, official letters, travel narratives, as well as works of a religious or literary nature. Legal texts that form part of a larger whole will also be included, for example, judicial proceedings, as will personal letters forming part of a larger narrative bundle.

### 3.3. Methodology

The texts of the corpus will be transcribed according to the guidelines established by the Seminary in Mackenzie (1997). HSMS’ semi-paleographic transcription system attempts to replicate, to the extent possible, various details related to the format and appearance of the text: folio and column number, original spelling, abbreviations and their resolution,

---

<sup>10</sup> The OPDT refers to the date that the text was originally produced while the SPDT refers to the date of the production of the specific manuscript copy or imprint. For example, internal evidence shows that the *Relación de la Jornada de Cíbola* was written sometime after the death of Joanna of Castile, so its OPDT is 1555 *a quo*; its SPDT is 1596, the date of the extant copy. See Faulhaber (1997–) regarding the dating of texts.

<sup>11</sup> These are the second, third, and fourth *relaciones*, published in 1522, 1523, and 1525, respectively (2CR, 3CR, and 4CR of the *Colonial Texts Corpus*; see Appendix).

<sup>12</sup> A noteworthy project that includes texts of this type is the *Corpus Diacrónico y Diatópico del Español* (CORDIAM), which deals exclusively with texts from Latin America. Archival documents are included in the subcorpus CORDIAM-Documentos.

upper- vs. lower-case letters, rubrics, glosses, headings, catch words, scribal errors and emendations, as well as editorial interventions (Gago Jover 2015). This allows the reader to reconstruct the format and appearance of the original text, ensuring philological integrity.

Contributors to the CCT project will edit their texts following philological best practices. Typically, the scholar will work from a digital facsimile and, when feasible, will correct the initial transcription by comparing it to the original text in the library or archive in which it is housed. The publication will follow the *Texts and Concordances* framework of the HSMS, with optional introduction, the transcribed text, the indices, and the concordances. These will be published in an open-access format on DLOST in the *Colonial Texts Corpus*. A link to digital images of the text will also be provided when available.

This methodology distinguishes the *Colonial Texts Corpus* from other corpora in important ways. First, all texts in the corpus are transcribed using the same editorial criteria. Other corpora, such as CORDE and Davies (2002–), incorporate texts that were edited using a wide variety of criteria—from paleographic transcriptions of a single manuscript or imprint to critical editions that reconstruct evidence from multiple extant versions of a text. Moreover, the present corpus eschews the inclusion of modern editions in which orthography is regularized, contrasting in this way with the two corpora cited above, as well as with CORDIAM.<sup>13</sup> The *Colonial Texts Corpus* provides access to a specific manuscript or imprint, with minimal editorial intervention. The corpus also employs uniform chronological criteria, giving preference to the SPDT over the OPDT. In contrast, other corpora prioritize the OPDT. In Davies (2002–), for example, fifteenth-century copies of Alfonsine texts are included in the database as thirteenth-century source material. The features of the *Colonial Texts Corpus* highlighted above allow researchers to extract reliable data with which to perform contrastive analyses, comparing apples to apples, as it were.<sup>14</sup>

---

<sup>13</sup> CORDE, for example, uses the modern edition by Hernández (1988) of Cortés' *Cartas de Relación*. CORDIAM makes use of modern editions in the subcorpus CORDIAM-Literatura, which includes chronicles as well as other textual types.

<sup>14</sup> See Gago Jover (2015: 10) for references to projects that use data from DLOST. To these can be added three lexical studies in progress whose data regarding indigenous loanwords, semantic extensions, and Arabisms largely derive from the *Colonial Texts Corpus*.

### 3.4. Current status of project

Preparation of the corpus began in 2017. After the parameters above had been determined, the principal investigators began to construct the beta version of the webpage. The initial nucleus of texts consisted of existing transcriptions from the Colonial Spanish Series of the HSMS which fit the established typological criteria. These were CIB, PMZ, and RVC (see Appendix). With these three, COL was included (this transcription was among the HSMS textual archives but had not been published), which brought the initial nucleus to four texts representing 200,799 tokens of data. The first texts that were added to the *Colonial Texts Corpus* were 2CR and VCC. When the project was launched in 2018,<sup>15</sup> the textual archive consisted of six texts (305,510 tokens). At present, the corpus consists of eleven texts (512,590 tokens) and will be continuously expanded. Collaborators in the project currently have eight additional texts in preparation, with another dozen in the planning stages.

### 3.5. Interface

As seen in Figure 1, the initial window displays the **navigation menu** ① and the name of the corpus of texts.

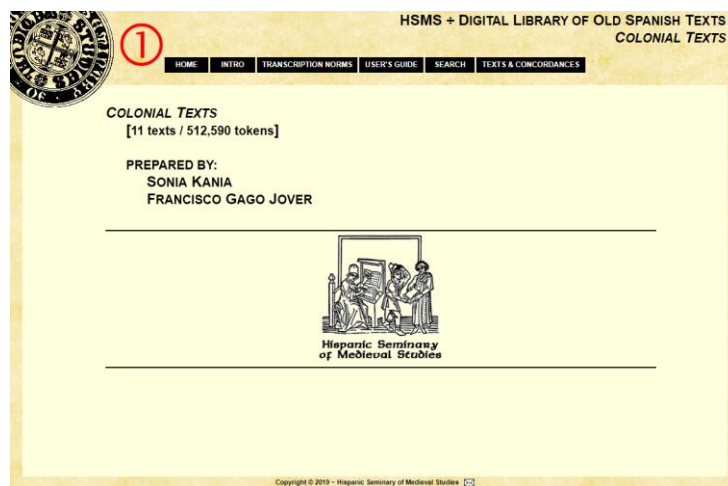


Figure 1: Initial window

The **navigation menu** ① provides quick access to:

- a. **Home:** The initial page, with general information on the texts and concordances.

<sup>15</sup> The project was formally launched at the *XI Congreso Internacional de Historia de la Lengua Española* in Lima, Peru, in August 2018.

- b. **Intro:** A brief overview of the CCT project.
- c. **Transcription norms:** A brief summary of the transcription coding used in the texts.
- d. **User's guide:** A brief explanation of the different parts of the interface.
- e. **Search:** The corpus search page.
- f. **Texts & concordances:** The interactive indexes, concordances, and texts described below.

Clicking on the **texts & concordances** button in the **navigation menu** ① brings up a list of all the works included ② (see Figure 2).



Year	Title	Library	HSMS id
1493	Carta a Luis de Santángel	New York: Public Library, 1423 Columbus	COL
1522	Segunda carta de relación	Providence: JCB Library, 1-512E B522 C8285	2CR
1523	Tercera carta de relación	Providence: JCB Library, 1-512E B523 C8284	3CR
1525	Cuarta carta de relación	Providence: JCB Library, 1-512E B523 C8286	4CR
1534	Vendición relación de la conquista del Perú	Providence: JCB Library, 1-512E B534 X81	VRP
1544	Relación de Francisco Vázquez de Coronado	Seville: ADL, Justicia 339, nº 1, ramo 1	RVC
1544	Apelación de Francisco Vázquez de Coronado	Seville: ADL, Justicia 339, nº 1, ramo 1	AVC
1552	Viajes de Cristóbal Colón	Madrid: Biblioteca Nacional, VITR6/7	VCC
1565	Relación del viaje de Pedro Menéndez de Avilés a la Florida	Seville: ADL, Patronato 19 B 17	RPM
1596	Relación de la jornada de Cibola	New York: Public Library, MiscCat 2570	CIB
1600-1602	Procedencia de mentes de Vicente Zañavier	Seville: ADL, Patronato 22 R4	PMZ

Figure 2: List of texts

As Figure 3 illustrates, clicking on one of the works opens up the **information menu** ③ of the selected item, which provides detailed information on the text and its concordances:

- a. **Title.**
- b. **Author.**
- c. **Translator.**
- d. **Specific Production Date.**
- e. **Original Production Date.**
- f. **Place of Production .**
- g. **Library:** Current location of the manuscript or imprint.
- h. **Printer:** Name of printer.
- i. **Transcribed by:** Name of the person(s) who transcribed the work.
- j. **Corrected by:** Name of the person(s) who corrected the transcription.
- k. **Lexical Studies:** Link to information in the *Lexical Studies of Medieval Spanish Texts* database (Dworkin and Gago Jover 2004–2018).



Figure 5: Wordlist, concordance, and text frames

At the left of the browser window is the **wordlist frame** ④, containing an **alphabetic list** of all words which are used in the source text. Clicking on a headword in the wordlist will make the **concordance frame** ⑤ scroll automatically to display all the instances of that headword, together with a line of context. The user can also select to see a **frequency** or a **reverse alphabetic** list. The **search box** allows the user to search for any word or combination of letters within each of the lists. The **concordance frame** ⑤ appears in the upper of the two large frames to the right of the wordlist. Beside each headword is a count of the number of times it occurs, and below it are all the occurrences, each in a line of context. To the right of each context line are the folio references. Clicking on a reference will make the **text frame** ⑥ scroll automatically to display the relevant part of the source text.

The text from which the concordance was made appears in the **text frame** ⑥, to the right of the wordlist. The scroll bars can be used to navigate in the text. To facilitate reading, the text is shown stripped of all transcription tags, with abbreviations resolved in italics, the combinations *c'*, *n~*, *s'*, and *z'* as *ç*, *ñ*, *σ*, and *ς*, respectively, and the *calderón* as ¶. In this way, the tagged transcription of the fragment in (1) of the *Carta a Luis de Santángel* (COL, fol. 1r) is shown with stripped tags (2):

- (1) puse nonbre la isla de santa maria de[ ]concepcion ala tercera ferrandina ala quarta la isla [isa]bella | ala qui<n>ta la Jsla Juana e asi a cada vna nonbre nuevo Quando yo lleg(\$u)[u]e ala Juana seg- | ui io la costa della al poniente yla falle tan grande q<ue> pense que seria tierra firme la proui<n>cia de | catayo y como no falle asi villas y luguares enla costa dela mar saluo pequen~as poblaciones [...]

- (2) puse nonbre la isla de santa maria de concepcion ala tercera ferrandina ala quarta la isla isabella | ala quinta la Jsla Juana e asi a cada vna nonbre nuevo Quando yo llegue ala Juana seg-ui | io la costa della al poniente yla falle tan grande *que* pense que seria tierra firme la prouincia de | catayo y como no falle asi villas y luguares enla costa dela mar saluo pequenas poblaciones [...]

### 3.6. Functionality

It is possible to search within a text or the entirety of the corpus. For the first type of search (within a single text) use the **text box** ⑦ as displayed in Figure 6. The search is performed in the selected index (alphabetic, frequency, or reverse); it is possible to anchor the search string to the beginning or the end of a word by using a bar (/), for example, /aceit, ndos/.

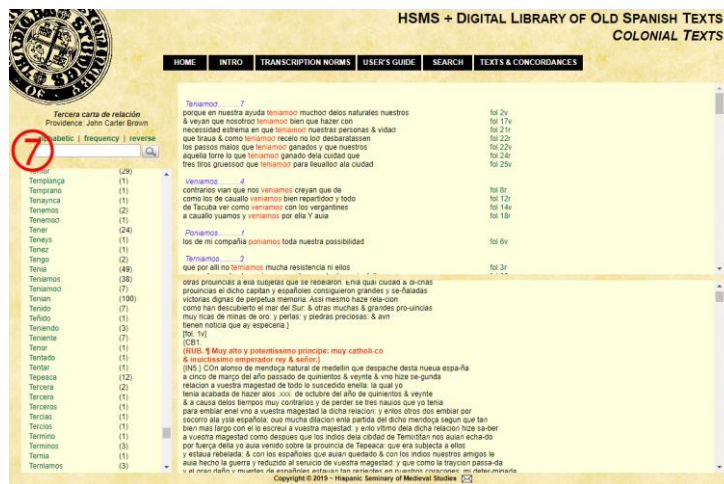


Figure 6: Search within a single text

As shown in Figure 7, to search the entirety of the corpus, click on the **search** button in the **navigation menu** ① to bring up the **search window** ⑧.

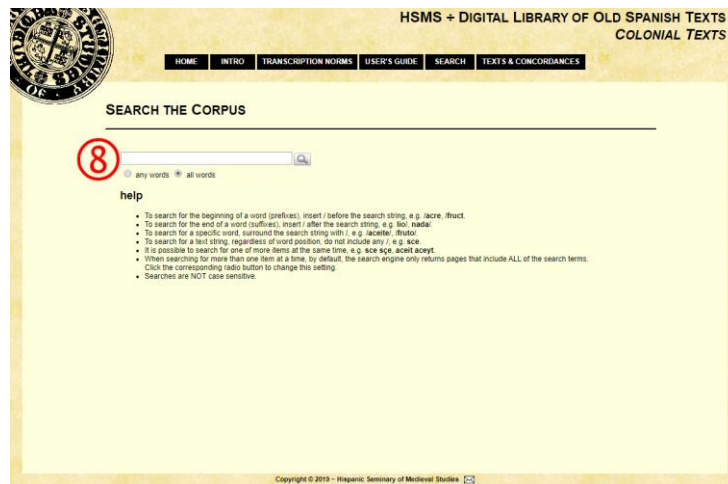


Figure 7: Search within the corpus

To search, type the search string in the text box.

- To search for the beginning of a word, insert / before the search string, e.g. **/acre**, **/fruct**.
- To search for the end of a word, insert / after the search string, e.g. **lio/**, **nada/**.
- To search for a specific word, surround the search string with /, e.g. **/aceite/**, **/fruto/**.
- To search for a text string, regardless of word position, do not include /, e.g. **sce**.
- It is possible to search for one or more items at the same time, e.g. **sce sçe**, **aceit aceyt**.
- When searching for more than one item at a time, by default, the search engine only returns pages that include all of the search terms. Click the corresponding radio button to change this setting.
- Searches are not case sensitive.

Searches are performed in the entirety of the corpus, and the search results page ⑨ shows all the texts in which the search string appears (see Figure 8). Clicking on the title brings up the concordances of the corresponding text.

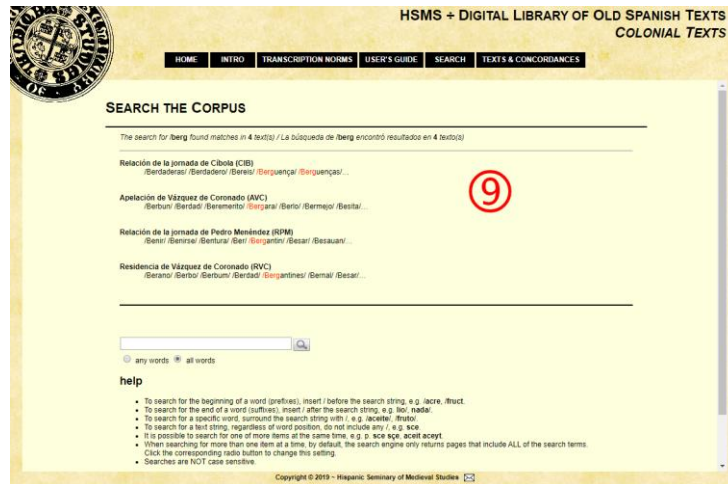


Figure 8: Search results page

Considering the nature of the *Colonial Texts Corpus*, more a repository of texts than a full-fledged linguistic corpus, and despite the limitations of the search engine, it is still possible to perform various types of queries. For example, looking at the distribution of forms such as *fijo(s)* / *hijo(s)*, one can observe that forms with initial *f*- only appear, in alternation with forms with *h*-, in 2CR, 3CR, and VCC, while the rest of the texts only present forms with *h*-. Looking at *fuesse* vs. *fuese* in all the texts, three of the eleven documents both *fuesse* and *fuese* (VCC, 4CR, PMZ), one (2CR) uses only *fuesse*, and four (AVC, RVC, CIB, RPM) use only *fuese*. By using the concordances, it is possible to study the use of any form in a specific text; for instance, in VCC, when studying the alternation of forms of the imperfect subjunctive in *-se* / *-sse*, a greater use of forms in *-se* (269 tokens: 94.1%) is observed compared to the forms in *-sse* (17 tokens: 5.9%). The preference for forms in *-se* can also be seen in the three verbs where both forms, *-se* / *-sse*, are present, as shown in (3).

- (3) dixese (4) / dixesse (1)  
       fuese (48) / fuesse (5)  
       llegase (6) / llegasse (2)

#### 4. CONCLUDING REMARKS

This article has presented a new subcorpus of the *Digital Library of Old Spanish Texts*, namely the *Colonial Texts Corpus*. The aim of the *Corpus of Colonial Texts* project is to provide open, interactive access to a corpus of texts that are transcribed using philologically rigorous criteria. This project therefore responds to the need for reliable primary sources related to colonial Latin America, a textual type that is under-represented

in existing electronic corpora. In this way, we will extend the reach of DLOST to a new group of users, i.e. scholars of colonial Latin America. Finally, this on-going project contributes to our broader goal of preserving the linguistic, cultural, and literary history of Spanish in the Americas.

#### REFERENCES

- Bethell, Leslie ed. 2002. *América Latina en la Época Colonial*. Vol. 1, *España y América de 1492 a 1808*. Barcelona: Crítica.
- CHARTA = *Corpus Hispánico y Americano en la Red: Textos Antiguos*. <http://www.corpuscharta.es>
- Company Company, Concepción. 2019. Voces e historia conceptual. Contribución a la construcción identitaria del español en América. Plenary talk given at the *Jornadas de Investigación El léxico americano en su historia: análisis y perspectivas de estudio*, Universidad de Querétaro, Querétaro, Mexico, October 2019.
- CORDE = Real Academia Española: Banco de datos (CORDE) en línea. *Corpus Diacrónico del Español*. <http://www.rae.es>
- CORDIAM = Academia Mexicana de la Lengua: *Corpus Diacrónico y Diatópico del Español*. <http://www.cordiam.org>
- Craddock, Jerry R. 2015. The Cíbola Project: Mission statement and staff. *UC Berkeley: Research Center for Romance Studies*. <https://escholarship.org/uc/item/3jt748vt> (7 January, 2020.)
- Craddock, Jerry R. and John H. R. Polt. 2008. An object lesson: Why we need good editions of the documents of the Hispanic Southwest. *UC Berkeley: Research Center for Romance Studies*. <https://escholarship.org/uc/item/6w33k9v5> (7 January, 2020.)
- Davies, Mark. 2002–. *Corpus del Español: 100 million words, 1200s–1900s*. <http://www.corpusdelespanol.org/hist-gen/>
- Delgado Gómez, Ángel ed. 1993. *Hernán Cortés, Cartas de Relación*. Madrid: Clásicos Castalia.
- Dworkin, Steven N. and Francisco Gago-Jover. 2004–2018. *Lexical Studies of Medieval Spanish Texts*. Hispanic Seminary of Medieval Studies, *La Corónica: A Journal of Medieval Hispanic Languages, Literatures, and Cultures*. <http://www.hispanicseminary.org/lsmst/index.htm> (7 January, 2020.)
- Faulhaber, Charles B. dir. 1997–. *BETA (Bibliografía Española de Textos Antiguos)*. The Bancroft Library. University of California, Berkeley. [http://vm136.lib.berkeley.edu/BANC/philobiblon/beta\\_en.html](http://vm136.lib.berkeley.edu/BANC/philobiblon/beta_en.html) (7 January, 2020.)
- Gago Jover, Francisco ed. 2011. *Digital Library of Old Spanish Texts*. Hispanic Seminary of Medieval Studies. <http://www.hispanicseminary.org/textconc-en.htm> (7 January, 2020.)
- Gago Jover, Francisco. 2015. La *Biblioteca Digital de Textos del Español Antiguo (BiDTEA)*. *Scriptum Digital* 4: 5–36.
- Gago Jover, Francisco and F. Javier Pueyo Mena. 2018a. El *Old Spanish Textual Archive*, diseño y desarrollo de un corpus de textos medievales: El corpus textual. *Cuadernos del Instituto Historia de la Lengua* 11: 165–209.
- Gago Jover, Francisco and F. Javier Pueyo Mena. 2018b. El *Old Spanish Textual Archive*, diseño y desarrollo de un corpus de textos medievales: Lematización y etiquetado gramatical. *Scriptum Digital* 7: 25–35.

- Hernández, Mario ed. 1988. *Hernán Cortés, Cartas de Relación*. Madrid: Historia 16.
- Lapesa, Rafael. 1981. *Historia de la Lengua Española* (ninth edition). Madrid: Gredos.
- Mackenzie, David. 1997. *A Manual of Manuscript Transcription for the Dictionary of the Old Spanish Language* (fifth edition by Ray Harris-Northall). Madison: Hispanic Seminary of Medieval Studies.
- Penny, Ralph. 2002. *A History of the Spanish Language* (second edition). Cambridge: Cambridge University Press.
- Torrens Álvarez, María Jesús. 2018. *Evolución e Historia de la Lengua Española* (second edition). Madrid: Arco Libros.
- Winship, George P. ed. and trans. 1896. *The Coronado Expedition, 1540–1542*. Bureau of Ethnology, Smithsonian Institution, Annual Report, 14. Washington, D.C.: Government Printing Office.

APPENDIX<sup>16</sup>

- COL** (1493): *Carta a Luis de Santángel*. New York: Public Library, \*KB + 1493 Columbus.
- 2CR** (1522): *Segunda Carta de Relación*. Providence: JCB Library, 1-SIZE B522 .C828c.
- 3CR** (1523): *Tercera Carta de Relación*. Providence: JCB Library, 1-SIZE B523 .C828ct.
- 4CR** (1525): *Cuarta Carta de Relación*. Providence: JCB Library, 1-SIZE B523 .C828r.
- VRP** (1534): *Verdadera Relación de la Conquista del Perú*. Providence: JCB Library, 1-SIZE B534 .X61.
- RVC** (1544–1545): *Residencia de Francisco Vázquez de Coronado*. Sevilla: AGI, Justicia 339, nº 1, ramo 1.
- AVC** (1544–1545): *Apelación de Francisco Vázquez de Coronado*. Sevilla: AGI, Justicia 339, nº 1, ramo 1.
- VCC** (1552): *Viajes de Cristóbal Colón*. Madrid: Biblioteca Nacional, VITR/6/7.
- RPM** (1565): *Relación del Viaje de Pedro Menéndez de Avilés a la Florida*. Sevilla: AGI, Patronato 19, R.17.
- CIB** (1596): *Relación de la Jornada de Cíbola*. New York: Public Library, MssCol 2570.
- PMZ** (1600–1602). *Probanza de Méritos de Vicente de Zaldívar*. Sevilla: AGI, Patronato 22, R.4.

*Corresponding author*

Sonia Kania  
 Department of Modern Languages  
 230 Hammond Hall, Box 19557  
 701 Planetarium Place  
 Arlington TX 76019  
 e-mail: skania@uta.edu

received: January 2019  
 accepted: March 2020

<sup>16</sup> Texts are listed in chronological order. Information provided includes three-character HSMS id, SPDT, title, library (preceded by city), and call number.

# Designing and building SCoPE<sup>2</sup>: A spoken corpus of Brazilian Portuguese and L2-English

Giovani Santos  
Mary Immaculate College/ Ireland

**Abstract** – This paper presents the process of designing and building a bilingual spoken corpus in order to pragmatically analyse oral L2-English discourse produced by a group of Brazilian university students living in Ireland. It discusses some of the decisions made, challenges faced, and considerations taken while designing a do-it-yourself corpus with a theoretical framework grounded in Corpus Pragmatics. The main objective is to share the lessons learned by examining the steps of designing and building SCoPE<sup>2</sup>, a bilingual spoken corpus, including the selection of participants, gathering data, and challenges in transcribing and coding spoken language with pragmatics in mind.

**Keywords** – bilingual corpus; spoken corpus; L2 corpus; corpus design; corpus construction; Corpus Pragmatics

## 1. INTRODUCTION

Answering many linguistic questions can be done considerably more easily using the increasing range of excellent and freely-available corpora. However, some specialised questions can only be answered by do-it-yourself (DIY) corpora. These are carefully designed and custom-built to assist in answering the questions and meeting the needs of specific research. The reasons to build a DIY corpus can be quite diverse, yet the analysis of such corpora can be very fruitful. Nevertheless, care and attention to corpus design is a must, as ill-conceived or poorly designed corpora can have, among other consequences, a serious negative influence on the research output.

This paper, resulting from a PhD study, presents the process of designing and building a bilingual spoken corpus in order to pragmatically analyse English discourse produced by a subgroup of native Brazilian speakers living in Ireland, with a specific focus on pragmatic markers (PMs).

The study is set within the fusion of Corpus Linguistics and Pragmatics, resulting in the emergence of a relatively new field termed Corpus Pragmatics (Aijmer and Rühlemann 2015). This is a blending which has evolved considerably over the last decade and which has proved to be a beneficial approach to investigating and understanding the usage of many linguistic features of real language-in-use evidenced by the rich expanding body of available studies.<sup>1</sup> Yet, most corpora are not designed for the considerations of pragmatics research and the analysis, categorisation and sometimes tagging of pragmatic features is conducted *post hoc*. It is in this context that there is need for careful consideration when designing a corpus for pragmatic research questions.

## 2. DIY: A BILINGUAL SPOKEN CORPUS FOR CORPUS-PRAGMATIC RESEARCH

The bilingual spoken corpus presented in this paper is targeted and designed with a context-specific research question in mind, namely to establish whether Brazilian university students in Ireland have accommodated particular features of the Irish English pragmatic system through their exposure to the native environment.

The main objective is that of investigating and describing the presence and function of PMs<sup>2</sup> in the participants' first (L1) and second (L2) languages. In addition, the study aims to compare and contrast the use of these features in the participants' L2 against their mother tongue (Portuguese) and the target language (English), so as to gain insights into the influence of both their L1 and the cultural immersion on their development of an L2.

Due to the specificity of the research question and aims of the study, a DIY corpus was necessitated, this being a bilingual corpus comprised of two sub-corpora, namely L1-corpus (Brazilian Portuguese) and L2-corpus (English as a Second Language), amounting to over 200,000 words across both languages. On the usefulness of small, carefully designed and targeted corpora, McCarthy and O'Keeffe (2010) observe that such corpora show evidence of being powerful tools to investigate and elucidate specific language use.

Another advantage of compiling a small corpus is the fact that the compiler and the analyst are often one-and-the-same: this gives an insider perspective to the analysis,

---

<sup>1</sup> For studies on different pragmatic phenomena through the lenses of Corpus Pragmatics see Romero-Trillo (2008) and Aijmer and Rühlemann (2015), among others.

<sup>2</sup> This paper subscribes to Fraser (1996) on the concept of PMs as a major umbrella under which many types of PMs constitute the class of "items which mark speakers' personal meanings, their organisational choices, attitudes and feelings" (Carter and McCarthy 2006: 207).

bolstering in the interpretation of the data. This close relationship between analyst and data, as well as language and context, makes small corpora a perfect fit for pragmatic studies. This advances the case for a Corpus Pragmatics (CP) framework (Rühlemann and Clancy 2018), which analyses the data both in a vertical quantitative manner (Corpus Linguistics), and in a horizontal qualitative manner (Pragmatics). This advantageous synergy makes CP a significant framework for the reliable, context-specific analysis of language use and language development.<sup>3</sup> Within the CP framework, the analyst can generate software-driven statistics while, concurrently, undertaking a detailed interpretation of the data considering the three major contexts in Pragmatics: situational, background knowledge and co-textual (Cutting 2002).

According to Jucker *et al.* (2018), CP is one of the three empirical methodological approaches for the analysis of pragmatic phenomena (the others being experimental pragmatics and observational pragmatics). What is more, PMs are perceived as one of the key areas of corpus-pragmatic research (Clancy and O’Keeffe 2015), which includes, but is not limited to, PMs across languages (see contributions in Aijmer and Simon-Vandenberg 2006), in L2 (Veiga 2016; Santos 2019), and in comparisons between native and non-native speakers (Aijmer 2004; Fung and Carter 2007).

The following sections describe the phases entailed in the designing and building of a bilingual DIY corpus, namely the *Spoken Corpus of Portuguese and English as a Second Language* (SCoPE<sup>2</sup>), which include: participant selection, and data collection and transcription. Owing to space constraints, this paper is limited to the description of the L2-corpus.

### 3. CORPUS DESIGN AND DESCRIPTION

There is common consensus that strict criteria to select the corpus type and participants are fundamental to answer the questions set in the research (Granger 2002; O’Keeffe *et al.* 2007; Adolphs and Knight 2010). A first step of considerable importance to take when designing a corpus is the matter of representativeness. Adolphs and Knight (2010) caution that it is the compiler’s responsibility to plan and predict any factor that may be a case for inconsistency and non-homogeneity in the corpus. This process involves not only the

---

<sup>3</sup> See Clancy and O’Keeffe 2015 for a critical review on research regarding key areas of CP.

selection of the participants, but also the type of material to be produced by them for data collection, as well as elements such as environment and technology.

Regarding this, the criteria for the current research to select the participants and achieve corpus representativeness are as follows:

- L2 context: University students in Ireland
- Mother tongue: Brazilian Portuguese
- Level of English Proficiency: C1-C2 CEFR (Common European Framework of Reference for Languages)

In addition to the criteria described above, the participants must be considered to be in the category of Successful Users of English (SUEs) which, according to Prodromou (2008), entails being able to engage accurately and fluently in different contexts with both L1- and L2- speakers of the target language, but does not necessarily equate to being native like.

The data in SCoPE<sup>2</sup> represents informal real language-in-use. However, Granger (2002) notes that collecting authentic data of L2 speakers might be a major challenge, as this kind of data is normally collected during task-based activities in class. To avoid the unnatural characteristic of classroom-elicited data, unscripted on-line language in use was collected with the participants' consent during casual meetings between the researcher and friends or fellow university students. The meetings took place in informal public and private settings according to the participants' convenience. The interactions were recorded using the iPhone Voice Memos app, which produced high audio quality (especially when recording dyadic interactions as interlocutors were near the microphone in contrast to one multiparty interaction with participants overlapping while conversing and occasionally moving in the room).

Table 1 details the corpus design in a data matrix.

<b>Number of participants</b>	17
<b>Gender</b>	Mixed (11 female and 5 male)
<b>Participants' age</b>	Adults (20-35 years old)
<b>Nationality</b>	Brazilian
<b>L2 Context</b>	University students in Ireland
<b>Level of L2 Proficiency</b>	C1-C2 CEFR; SUE (Prodromou 2008)
<b>Type of data</b>	Audio recordings of unscripted informal conversations
<b>Type of interaction</b>	13 dyadic (participant + researcher) and 1 multiparty (3 participants + researcher)
<b>Average duration</b>	30 minutes in each language (L1 and L2)

Table 1: Data matrix

These informal conversations have a main topic in common. The participants are mostly discussing their experiences and perceptions of travelling in Brazil and all around the world, though many other topics (such as personal relationships, future goals, etc.) may arise in the conversations due to their natural and unscripted nature. When both contributions (L1 and L2) from a participant were recorded on the same occasion, the researcher greeted them in English and used the L2 as an icebreaker before the actual recording, rather than their common L1. The L2 was thus recorded first, followed by the L1.

It is important to note that the L2 sub-corpus described in this paper is not a learner corpus but, in fact, a corpus of proficient L2 language. This is in view of the fact that all participants in this research have successfully completed their English language programmes, have achieved an internationally recognised certificate of either advanced or proficient English, and communicate efficiently within an international environment both with L1- and L2-speakers from different first-language backgrounds, be it at personal, professional or cultural levels. This perspective of an L2 corpus, rather than a learner corpus, accords with that of Prodromou's (2003) as well as with his search and categorisation of SUEs to build such a corpus.

#### 4. TRANSCRIPTION CONVENTION

The collected data comprises informal real spoken language-in-use, and thus its transcription is required for further corpus-based analysis. As noted by Kirk and Andersen (2016: 295), transcriptions are a representation of spoken language which is subjected to a process of "selection, abstraction and omission." It is, therefore, the transcriber's responsibility to ensure that the transcription is as truthful to its spoken version as

possible, though the amount of detail, marking and coding will always depend on the specific research needs (Adolphs and Knight 2010). This section presents the rationale behind the choices for and needs of the transcription of a corpus designed to analyse PMs within a context of L2 development.

#### *4.1. Transcribing spoken language*

Due to the multimodal nature of spoken language, the practice of transcription can pose a real challenge when it comes to deciding what is to be included from the array of detailed layers which can be extracted from an interaction (Adolphs and Knight 2010). Spoken interaction is not simply comprised of utterances alone, but these function alongside several extralinguistic features such as tone, rhythm, laugh, eye gaze, and body movement, in order for a speaker to convey a message and their listener(s) to infer the intentional meaning in the speaker's proposition. As Adolphs and Knight (2010: 44) aptly put it, spoken interaction features "a careful interplay between textual, prosodic, gestural and environmental elements in the construction of meaning."

Some researchers may wish to build sound-text aligned corpora, providing resources to undertake studies focused not only on the structure of the spoken language, but also on its prosodic features (see contributions in Cresti and Moneglia 2005). Others may focus on a wide range of pragmatic phenomena, thus annotating their corpora in order to investigate pragmatic intent (Kirk 2016). For those who are interested in the understanding of talk in interaction, Conversation Analysis provides a thorough transcription convention in order to analyse interactive features such as overlapping, prosody and non-verbal elements (Liddicoat 2007).

Adolphs and Knight (2010) flag the individuality of each research project and the importance of identifying with precision the purpose of the study prior to selecting the type of transcription for the study in question, noting that "[i]t is advisable to identify the spoken features of interest at the outset, and to tailor the focus of the transcription accordingly" (2010: 44). With that in mind, SCoPE<sup>2</sup> was built to analyse PMs in L2, with a view towards the influence which the speakers' L1 and their exposure to the target language may have on their production of such linguistic features. This means that a detailed annotation of different pragmatic phenomena was not necessitated, as the type of corpus methodology adopted was that of a form-to-function, rather than function-to-form

(O’Keeffe *et al.* 2019). In other words, frequency and keyword lists, as well as previous studies on the linguistic feature under investigation, are used as starting points, thus a broad transcription with minimum annotation was sufficient.

#### 4.2. Transcription convention: SCoPE<sup>2</sup>

The transcription convention used in the structure of SCoPE<sup>2</sup> (e.g. codes, tags, and punctuation of both sub-corpora) was adapted from that employed in the transcription of the *Limerick Corpus of Irish English* (LCIE; Farr *et al.* 2004) which, in turn, was based on the *Cambridge and Nottingham Corpus of Discourse in English* (CANCODE; McCarthy 1998). This was a natural choice since the LCIE is used as a reference corpus in the study for which SCoPE<sup>2</sup> was built, as it represents the variant of English to which the participants have been exposed. Similarly, the transcription convention for the linguistic material of the L2-corpus of SCoPE<sup>2</sup> (e.g. filled pauses, backchanneling, contractions, etc.) was adapted from that of the LCIE. However, the transcription convention for the linguistic material of the L1-corpus was adapted from that of the C-ORAL-BRASIL (Mello *et al.* 2012), chosen as a model due to its thorough description of decisions and rationale when working with spoken Brazilian Portuguese.

Although some adaptations were needed, and a parallel transcription was not necessarily required between the LCIE and SCoPE<sup>2</sup> (due to different research foci), it is useful to try to bring a corpus as close as possible to its reference when it comes to the transcription, as this will facilitate future data reading and analysis. Most of the codes used in the LCIE transcription maintained their original functions in both SCoPE<sup>2</sup> sub-corpora in their original functions, while others were slightly adapted within their existing functions. It is also important to note that, despite obvious variations regarding the transcription of linguistic material between two languages (e.g. English backchannelling *mmhm* versus Brazilian Portuguese backchannelling *hum hum*; English language contractions such as *it’s* versus Brazilian Portuguese apheresis in the conjugation of the verb *estar* ‘to be’ in nearly all of its conjugations, e.g. *tá, tava, tô* etc.), the two parts of SCoPE<sup>2</sup> are fully comparable in their structure and codes. The remainder of this section is devoted to describe these adaptations and their justifications.

#### 4.2.1. General codes

For the identification of speakers, each participant is given a number according to their gender and order of file transcription, the odd numbers being males and the even numbers females.

To ensure the anonymity of anyone mentioned throughout the conversations, their names are replaced by either the speakers' own identification numbers (e.g. \$2 if speaker two mentions her own name) or a name that reflects the culture and/or language of the names that are mentioned. The string (1) gives an example where speaker \$2 is talking about her trip to Cuba when she mentions her hosts' names, who are both Cuban, and had their names replaced by names which are also Spanish sounding.

- (1) <\$2> Yeah er the the family we stayed with the first one <\$E> **pause** </\$E> **Juan** and **Rosa** uhm she was a psychologist and she used to get paid uhm eighteen euros a month.

As shown in (1), extra linguistic features are transcribed within the <\$E> </\$E> codes. Not only do they include significant extra information that happens during the interaction (e.g. pause, laugh, etc.), but they also include contextual and/or cultural background information (e.g. explanations of word play that draw on common cultural understandings).

As far as extra linguistic information is concerned, two features –laughs and pauses– which are natural to spoken interaction required further description. Laughing is thus divided into three categories: *laugh*, which refers to a loud and free expression of amusement; *chuckle*, a shorter and inner type of laugh; and *giggle*, describing an even shorter and lighter type of laugh. In the L1-corpus, these pieces of information are transcribed as *risada*, *risos* and *risadinha*, respectively.

Although pauses, at first glance, seem to be quite a straightforward feature to transcribe using the pairs short/long and filled/unfilled as a reference model, a system based on specific criteria needed to be developed to ensure consistency and ease of reading throughout the data. Having said that, it is important to note that filled pauses are only transcribed one way in each language, namely *uhm* for English and *eh* for Portuguese, rather than trying to develop an infinite and exhaustive list of different sounds produced by speakers when filling a pause (especially in L2 production). While *uhm* and *eh* may not cover all sounds produced by the participants either in English or Portuguese, they represent the act of filling space in language use while speaking. The choice for *uhm*

and *eh* over many other graphic forms of filled pauses is due to the unlikelihood of their being similar to any written word in either languages in the data, thus avoiding any possible misinterpretation when reading the texts.

Unfilled pause codes, on the other hand, serve to mark where pauses take place, either within an utterance or after it. Examples (2) and (3) present speakers' turns where it is possible to see two different functions in the use of such a feature. In (2), speaker \$2 makes use of a short pause to give emphasis to the adverb *actually* when asking speaker \$1 if he had been to Paraguay, whereas in example (3) a pause is employed to allow time to the speaker in order to restructure a sentence. These are examples of short pauses (up to three seconds). Longer pauses are coded according to their length, e.g. <\$E> pause of six seconds </\$E>.

- (2) <\$2> +but <\$E1> chuckle </\$E1> did you go there <\$E> **pause** </\$E> actually?
- (3) <\$1> I think there is always <\$E> **pause** </\$E> there are always both sides you know+

A break between two or more utterances produced within a speaker's turn with no interruption is marked with a full stop. A full stop, therefore, marks where one complete utterance finishes and another starts, as well as where a speaker's turn ends, as seen in example (4). Complete utterance refers to a complete thought, which can be either a full sentence or simply a phrase, and can occur as a full turn or within a turn. Alternatively, when a speaker's turn is not finished but still slightly interrupted by a short answer or backchannelling, the plus symbol (+) is used to mark such a phenomenon of speech continuity (see (4)).

- (4) <\$3> I don't know it's a tough question <\$E> pause </\$E> now I'll continue the list and **then**+  
 <\$1> Oh please yeah **yeah**.  
 <\$3> +**and** then I think about the favourite **place**. Okay <\$E> pause </\$E> after Lithuania I went to Tunisia+

Following the same rationale behind the use of a full stop to mark the end of an utterance, the equal symbol (=) is employed to identify where incompletions take place in the conversations. At a lexical level, an incomplete word can be either followed by its full form, as in the truncation seen in (5) where speaker \$5 is talking about free walking tours, or followed by another different word altogether, as in (6), where speaker \$3 restructures his sentence, changing its aspect.

- (5) <\$5> But you but you kinda feel bad when they like they don't charge anything but **the= they** ask for tips+
- (6) <\$3> +if you compare like to England. I **I don= I've** never been to England to be honest yeah no.

Alternatively, at a speaking-turn level, a word or an utterance may be simply interrupted by another speaker, as shown in (7).

- (7) <\$4> Yeah. **So=**  
 <\$1> How old are you?  
 <\$4> I'm twenty-four.

#### 4.2.2. Transcription issues regarding PMs

Considering that SCoPE<sup>2</sup> was designed with the particular research purpose of analysing PMs, two issues required particular attention during the transcription process, namely overlapping and ambiguity.

Although overlapping is an important feature in spoken interaction, the research for which SCoPE<sup>2</sup> has been designed is not grounded on Conversation Analysis, where features such as overlapping, as well as pauses, intonation, repair, etc. must be carefully marked. Therefore, overlapping was not marked because, besides being a time-consuming task, that would result in an over-coded text rather than aiding in the analysis.

Nevertheless, overlaps still pose transcription challenges in terms of how to maintain a natural and truthful flow of spoken language without marking where the overlaps take place in the conversation. An attempt to try to overcome some of these challenges is to follow the natural sequence of events in the conversation. Extract (8a) demonstrates where the overlaps take place (marked with the <\$O> </\$O> codes), while extract (8b) presents the same piece with the overlaps replaced by interruption marks (+ symbol) breaking the interaction in a natural sequence of events. In this interaction, speaker \$2 is explaining to speaker \$1 where the state of Paraná is located in Brazil.

- (8a) <\$1> So you are from the last state like the last one in the south.  
 <\$2> Mm <\$O1> **uhm** </\$O1>.  
 <\$1> <\$O1> **Paraná** </\$O1> <\$E> trying to locate Paraná in an imaginary map in the air </\$E>.  
 <\$2> No no no it's </\$O1> **Paraná** </\$O1> Santa Catarina and Rio Grande do Sul.  
 <\$1> <\$O1> **Rio Grande do Sul** </\$O1>.

- (8b) <\$1> So you are from the last state like the last one in the south.

- <\$2> Mm **uhm**.  
 <\$1> **Paraná** <\$E> trying to locate Paraná in an imaginary map in the air  
 </\$E>.  
 <\$2> No no no it's **Paraná**+  
 <\$1> Rio Grande do Sul.  
 <\$2> +**Santa Catarina** and Rio Grande do Sul.

Special attention, however, had to be given to the position of PMs in the utterances when breaking overlaps into turn continuity. Occurrences of minimal response tokens (e.g. *yeah, okay, mmhm, mm*) overlapping PMs, for example, were transcribed having PMs as priorities due to their multi-functionality. Considering the importance of the relationship between position and function when it comes to PMs, two linguistic phenomena were assessed: the co-occurrence of PMs and their prosodic aspects (e.g. short pause before or after the PM, rhythm, stress, pitch movement when delivering the PM). Extract (9a) illustrates a case where *you know* and *like* co-occur when speakers \$1 and \$2 are talking about how difficult flying to Cuba is. Speaker \$1 overlaps \$2 in between the pair of PMs:

- (9a) <\$2> Mm yeah so going to Cuba is sort of tricky <\$O1> **you know** </\$O1>  
**like** there are some since of the the embargo like the the United States  
 embargo in Cuba <\$O1> like </\$O1> it's sort of like not every uhm=  
 <\$1> <\$O1> **Mmhm** </\$O1>. <\$O1> Ah okay </\$O1>. It used to. Not  
 anymore.

To avoid a break between the original co-occurrence of two PMs which, in turn, avoids misinterpretation of their functions, the response token (RT) was placed in the next line after the PMs, as illustrated in (9b):

- (9b) <\$2> Mm yeah so going to Cuba is sort of tricky **you know like**+  
 <\$1> **Mmhm**.  
 <\$2> +there are some since of the the embargo like the the United States  
 embargo in Cuba like+  
 <\$1> Ah okay.  
 <\$2> +it's sort of like not every uhm=  
 <\$1> It used to. Not anymore.

Likewise, PMs were kept in their final or initial positions based on their prosodic aspects or way in which they were originally uttered by the speaker. This is illustrated by extracts (10a–b), where speaker \$2 is talking about her motivation to go to Cuba.

- (10a) <\$2> +already and uhm I had a job in Brazil I I graduated from law school  
 and got a job as a legal adviser and I knew I was going to move to Europe to  
 do a Masters and I I said “oh I’m going to be poor <\$E1> laugh </\$E1> in a  
 while” <\$E> pause </\$E> like not poor like <\$E> pause </\$E> but “I’m  
 going to be a student again <\$O1> in a while </\$O1> so now that I’m like  
 getting <\$O1> **paid**” </\$O1> **you know** “if I don’t go to Cuba now”+

<\$1> <\$O1> Yeah </\$O1>. <\$O1> **Mmhm** </\$O1>.

- (10b) <\$2> +already and uhm I had a job in Brazil I I graduated from law school and got a job as a legal adviser and I knew I was going to move to Europe to do a Masters and I I said “oh I’m going to be poor <\$E1> laugh </\$E1> in a while” <\$E> pause </\$E> like not poor like <\$E> pause </\$E> but “I’m going to be a student again in a while+  
 <\$1> Yeah.  
 <\$2> +so now that I’m getting **paid**” **you know**+  
 <\$1> **Mmhm**.  
 <\$2> +“if I don’t go to Cuba now”+

As seen in (10a), the overlap takes place before *you know*. However, moving the position of the PM from after the word *paid* to before the word *if* would result in a change of focus on the content being delivered. By listening to the recording, the researcher was able to ascertain the position of the PM within the speaker’s turn by taking into consideration prosodic cues. In this specific case, the word *paid* is linked to *you know* in the rhythm of the speech delivered, and the PM *you know* is then followed by a short pause before the speaker continues with her speech. Therefore, again, the position of the PM had priority over the position of the RT, as illustrated in (10b).

A case of potential ambiguity for the interpretation of PMs surfaced during the transcription process. The transcription of SCoPE<sup>2</sup> has only three pieces of punctuation (following the convention of the LCIE), the question mark being for questions, inverted commas for quotes and the full stop to close complete utterances. Consider examples (11) to (13) below:

(11) <\$1> You don’t **like** dogs?

(12) <\$2> I think I’ve been **like** to twenty twenty and a few countries.

(13a) <\$3> Uhm it’s just **like** a big city.

In (11), it is clear that *like* functions as a verb, while in (12) it functions as a PM. However, it is difficult to assign either a grammatic or a pragmatic function to *like*, as in (13a) it could be either of the options. Once again, during the transcription, prosody played an important role when differentiating between grammar and pragmatics. If the sentence in (13a) were uttered straight on with no hesitation, then it would be a case of *like* as a preposition. If the example were uttered with a slight pause between *like* and a *big city*, though, then this would be a case of PM, thus being transcribed with the <.> code as illustrated in (13b).

(13b) <\$3> Uhm it's just **like** <.> a big city.

#### 4.2.3. Customised new codes

Two more codes were included in the transcription convention in order to attend to the research design of SCoPE<sup>2</sup>, namely code-switching (<\$CS> </\$CS>) and language error/correction (<\$X> | </\$X>).

The first one, <\$CS> </\$CS>, refers to any code-switching occurring in either sub-corpus. The second one, <\$X> | </\$X>, is employed when a language error or mistake needs correction to avoid misinterpretation or confusion. The interpretation and correction of such mistakes were facilitated by the fact that the transcriber shares with the participants the same L1. The use of an error code was not a surprise due to the nature of the English sub-corpus, where the participants, although at proficient level of competency in English, were still developing their L2, and thus were not expected to be totally error-free.

Extract (14) illustrates a case where an entire string is composed of a series of mistakes which confuse the message. Without correction, it is unlikely that it would be understood by either an L1-user of English or L2-users with different language backgrounds. Speaker \$4 is talking about her willingness to come to Ireland for a year of study-abroad experience and her decision to break up with her boyfriend if he had decided not to come with her. Everything added after the vertical bar symbol (|) is the correction offered by the researcher:

- (14) <\$4> +“if you're not we are going to <\$X> **finish** | **break up** </\$X> right away here because I'm not going to <\$X> **sustain** | **maintain** </\$X> a relationship <\$X> **so far**| **over such a long distance** </\$X>+

## 5. CONSIDERATIONS AND CONCLUSION

As seen throughout this paper, the process of designing and building a totally new spoken corpus entails a series of steps that progress from the research questions to the transcription of the data. The type of research one is conducting plays a significant role in this process and so does the nature of spoken language itself, which can, at times, be incoherent and challenging, and thus pose many difficulties when giving it a written form.

To achieve a satisfactory standard in the data, a meticulous review of the work must be conducted with trials and plentiful patience. A number of pilot transcriptions must be undertaken and reviewed before a final and appropriate model can be achieved. The model described in this paper has been reviewed during its development in different ways:

- by reading a sample without listening to the audio in order to check if the transcription of spoken language is coming across as a true representation of the recorded interaction;
- by reading a sample while listening to its audio in order to double check the impressions during the first reading;
- by testing the tags/codes with CL software;
- after attending conferences and meetings on the field in order to gain feedback on the proposed coding system;
- by testing the convention against existing ones.

Another factor to consider when dealing with spoken language is that data collection may affect the transcription process and therefore special attention must be paid to elements such as background noises, the type of recording devices used and the number of speakers in the recorded conversation. For example, considering the natural and sometimes chaotic nature of spoken language, a multi-party conversation may actually be quite fruitless if not well managed, since the level of overlapping can be higher in comparison to a dyadic interaction, and the speakers may go off topic by interacting with different parties at the same time. All this natural messiness and diversity in language might be valuable for Conversation Analysis, but poses many challenges in collecting and maintaining a certain level of quality and truthfulness to spoken language.

Delaying the transcription of interactions can also cause significant problems. As previously mentioned, being the one-and-same data collector and transcriber contributes valuable insights and depth to the conversations and interactions recorded and transcribed. However, it is good practice and strongly recommended that transcriptions be carried out in the shortest time possible after the material has been collected, especially as many of the resources used to aid in online communication, such as body language and facial expressions, are still fresh in mind and, therefore, most helpful in aiding the transcription process and resolving any possible uncertainties or ambiguities that may arise.

In conclusion, though laborious, the DIY is a worthwhile type of corpus. The entire process requires one to reflect on the research questions and contemplate the analysis phase of the research. In addition, although these corpora may often be small in size, a wide range of linguistic features can be extracted for detailed analysis, making such corpora a rich source for corpus studies.

Each DIY corpus compiler may have their own rationale and theories to ground their design and construction. This paper has provided some practical experience and reflections on how to design and build a bilingual corpus to conduct language analysis within a Corpus Pragmatics framework. It is hoped that this work may open more discussions on how to achieve truthfulness and accuracy as close as possible to spoken language. Moreover, it is hoped that this may be a guiding resource for those aiming to venture on a DIY corpus journey for the first time, doing so in a critical way.

#### 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.
- Aijmer, Karin. 2004. Pragmatic markers in spoken interlanguage. *Nordic Journal of English Studies* 3/1: 173–190.
- Aijmer, Karin and Anne-Marie Simon-Vandenberghe eds. 2006. *Pragmatic Markers in Contrast*. London: Elsevier.
- Aijmer, Karin and Christoph Rühlemann eds. 2015. *Corpus Pragmatics: A Handbook*. Cambridge: Cambridge University Press.
- Carter, Ronald and Michael McCarthy. 2006. *Cambridge Grammar of English: A Comprehensive Guide*. Cambridge: Cambridge University Press.
- Clancy, Brian and Anne O'Keeffe. 2015. Pragmatics. In Douglas Biber and Randi Reppen eds. *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press, 235–251.
- Cresti, Emanuela and Massimo Moneglia eds. 2005. *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: John Benjamins.
- Cutting, Joan. 2002. *Pragmatics and Discourse: A Resource Book for Students*. London: Routledge.
- Farr, Fiona, Brona Murphy and Anne O'Keeffe. 2004. The Limerick corpus of Irish English: Design, description and application. *Teanga* 21: 5–29.
- Fraser, Bruce. 1996. Pragmatic markers. *Pragmatics* 6/2: 167–190.
- Fung, Loretta and Ronald Carter. 2007. Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics* 28/3: 410–439.
- Granger, Sylviane. 2002. A bird's-eye view of learner corpus research. In Granger, Sylviane, Joseph Hung and Stephanie Petch-Tyson eds. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins, 3–33.

- Jucker, Andreas H., Klaus P. Schneider and Wolfram Bublitz. 2018. Preface to *Methods in Pragmatics*. In Andreas H. Jucker, Klaus P. Schneider and Wolfram Bublitz eds. *Methods in Pragmatics*. Berlin: Mouton De Gruyter, x.
- Kirk, John M. 2016. The pragmatic annotation scheme of the SPICE-Ireland corpus. *International Journal of Corpus Linguistics* 21/3: 299–322.
- Kirk, John M. and Gisle Andersen. 2016. Compilation, transcription, markup and annotation of spoken corpora. *International Journal of Corpus Linguistics* 21/3: 291–298.
- Liddicoat, Anthony J. 2007. *An Introduction to Conversation Analysis*. New York: Continuum.
- McCarthy, Michael. 1998. *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.
- McCarthy, Michael and Anne O’Keeffe. 2010. Historical perspective: What are corpora and how have they evolved? In Anne O’Keeffe and Michael McCarthy eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 3–13.
- Mello, Heliana, Tommaso Raso, Maryualê M. Mittmann, Heloísa P. Vale and Priscila O. Côrtes. 2012. Transcrição e segmentação prosódica do corpus C-ORAL-BRASIL: Critérios de implementação e validação. In Tommaso Raso and Heliana Mello eds. *C-ORAL-BRASIL I: Corpus de Referência do Português Falado Informal*. Belo Horizonte: Editora UFMG, 125–176.
- O’Keeffe, Anne, Brian Clancy and Svenja Adolphs. 2019. *Introducing Pragmatics in Use* (second edition). London: Routledge.
- O’Keeffe, Anne, Michael McCarthy and Ronald Carter. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.
- Prodromou, Luke. 2003. In search of the Successful User of English: How a corpus of non-native language could impact on EFL teaching. *Modern English Teacher* 12/2: 5–14.
- Prodromou, Luke. 2008. *English as a Lingua Franca: A Corpus-based Analysis*. London: Continuum.
- Romero-Trillo, Jesús ed. 2008. *Pragmatics and Corpus Linguistics: A Mutualistic Entente*. Berlin: Mouton de Gruyter.
- Rühlemann, Christoph and Brian Clancy. 2018. Corpus linguistics and pragmatics. In Cornelia Ilie and Neal R. Norrick eds. 2018. *Pragmatics and its Interfaces*. Amsterdam: John Benjamins.
- Santos, Giovani. 2019. Second language pragmatics: A corpus-based study of the pragmatic marker *like*. *Letrônica* 12/4: 1–16.
- Veiga, Nancy V. 2016. Discourse markers in CEDEL2 and SPLLOC corpora of learner Spanish: Analysis of some lexical-pragmatic failures. In Margarita Alonso-Ramos ed. *Spanish Learner Corpus Research: Current Trends and Future Perspectives*. Amsterdam: John Benjamins, 267–297.

*Corresponding author*

Giovani Santos  
 Mary Immaculate College  
 Department of English Language & Literature  
 South Circular Road  
 V94 VN26, Limerick  
 Ireland  
 e-mail: giovani.santos@mic.ul.ie

received: February 2020

accepted: March 2020

# Brazilian cultural markers in translation: A model for a corpus-based glossary

Rozane Rebechi<sup>a</sup> - Stella Tagnin<sup>b</sup>  
Universidade Federal do Rio Grande do Sul<sup>a</sup> / Brazil  
Universidade de São Paulo<sup>b</sup> / Brazil

**Abstract** – Translations in the Brazilian culinary domain are often characterized by the use of inaccurate equivalents, a lack of fluency, and adaptations that lead to a mischaracterization of cultural references. This is due to a lack of reliable reference materials in that area which usually only offer a translation, without any context or explanation. To address these issues, this paper draws upon a corpus-informed methodology to devise a three-level entry – term/equivalent, appositive explanation and encyclopedic information – for Brazilian cooking terms in a Portuguese-English glossary aimed at translators and writers of culinary texts.

**Keywords** – Brazilian cooking terminology; cultural markers; corpus linguistics

## 1. INTRODUCTION

*Why should a word in a recipe be less important than a word in a novel? One can lead to physical indigestion, the other to mental.* (Barnes 2003: 7)

As well as being one of the fundamental elements of human existence, food is a distinctive cultural constituent of every nation. Despite being a popular topic, it is rarely regarded as a theme worthy of serious academic study (Brien 2007). Fortunately, this picture has changed over the past few years, with a considerable amount of literature on the culinary arts being recently published. Academic research has resulted in monographs, book chapters and papers (see, for example, Gerhardt *et al.* 2013; Jurafsky 2014; Temmerman and Dubois 2017; Tigner and Carruth 2018), and the connection between food and translation has given rise to scholarly events, such as the *International Conference on Food and Culture in Translation* (FaCT), held in Italy in 2014 and 2016. Nevertheless, as a result of the long neglect of the culinary domain at academic level (see Capatti and Montanari 1999), few reference materials have been published in the area, at least as far as the Portuguese-English language pair is concerned, and this paucity has had negative consequences for translation studies.

Food items “wander around the globe” (Gerhardt 2013: 17) in such a way that the cuisine of any given nation can be accessed by anyone and anywhere. Brazilian cooking is no exception. Driven by both international sports events held in the country –2014 FIFA World Cup and 2016 Olympic Games– and the worldwide renown of chefs who privilege local products, Brazilian cooking has received special attention from foreign audiences, judging by the number of Brazilian cookbooks published in English.<sup>1</sup> Unfortunately, these publications also reveal misunderstandings regarding Brazilian cuisine.

A melting pot basically influenced by Portuguese colonizers, Native Indians, and African slaves, typical Brazilian cooking is very rich in native, endemic and exotic ingredients and local preparations. Nevertheless, it is commonly reduced to a few items in general cooking dictionaries and bilingual glossaries, which often fail to define and/or translate them accurately and consistently. Through an examination of the few existing bilingual cookery dictionaries and glossaries, in addition to general monolingual and bilingual dictionaries available on the Brazilian market, Rebechi (2015a, 2015b) demonstrated that most of them are ineffective in providing equivalents and/or appropriate definitions for many typical Brazilian products. Contextualized examples, which could enhance comprehension, are also absent. As a consequence, texts related to Brazilian cooking often display mistranslations, inaccurate definitions of terms and substitution of ingredients, generating products and dishes that are not representative of our national cuisine. We believe that these problems might have been easily addressed if reliable lexicographical references were available.

One of the greatest challenges of compiling a reference material in the area is ensuring that the distinctive cultural characteristics of Brazilian cooking are maintained. Thus, we believe that a representative reference work aimed at translators in the area should not only provide equivalents for terms –whenever they have an equivalent– but also offer other key information which could be used in the translation. The main purpose of this article is to offer a model of a glossary entry which has been customized specifically for Brazilian culinary items, based on a corpus of authentic texts containing cooking recipes. The entries are meant to provide translators and writers of culinary texts

---

<sup>1</sup> The combination of the search words *Brazilian* and *cooking* resulted in 240 titles available on the *Amazon* online store. (9 January, 2020.)

with terms –along with their English equivalents–, appositive explanations, and encyclopedic information about typical Brazilian food items. Examples of use and phraseological units are included as additional data. To achieve our goal, we rely on Corpus Linguistics (CL) procedures which allow for an analysis of the term in context (see Pearson 1998).

## 2. CULTURAL MARKERS IN TRANSLATION

According to Newmark (1988: 94), culture refers to “the way of life and its manifestations that are peculiar to a community that uses a particular language and its means of expression.” He emphasizes that words which are characteristic of a certain culture (what we call cultural markers) will pose problems to the translator, unless source and target languages overlap. Cultural markers are here understood as textual, lexical, and discursive elements relating to specific cultures (Zavaglia *et al.* 2011), which can be found in any text type, either general, literary or specialized. Also called identity markers, they refer to various elements used to display preferences towards other cultures. As an important expression of a culture, food-related items demand a number of decisions when they are translated from one culture to another (Newmark 1988).

Still according to Newmark (1988: 97), “[f]ood is for many the most sensitive and important expression of national culture.” Hence, when shared with other cultures, food items are subject to a number of procedures, depending on the purpose of the translation. Regarding the translation of Brazilian cultural markers, we observe recurrent choices for functional equivalents (see Nord 2001, 2012). By rendering cultural words with culture-free words, the translator or writer neutralizes or generalizes a term (Newmark 1988). This procedure may result in a successful strategy regarding the translation of recipes when the aim is to guide a cook toward the appropriate preparation of a dish. For example, replacing buttermilk with a mixture of yogurt and milk will not affect the final result dramatically. On the other hand, using a different kind of bean, instead of the *feijão fradinho* ‘black-eyed peas’ to prepare *acarajé* (spoonful-sized fritter made from puréed black-eyed peas seasoned with salt and onion, deep-fried in *dendê* oil)<sup>2</sup> will certainly render a quite different type of bean fritter. Besides, what motivates a foreigner to choose

---

<sup>2</sup> *Dendê* oil is a thick, dense, reddish oil with a delectable flavor and aroma used as an ingredient in Bahian dishes.

a Brazilian recipe is probably the exoticism of our typical foods. Similarly, when Brazil's national drink *cachaça* is rendered as 'crude/sugarcane brandy', 'spirit(s)' or '(sugarcane) rum', it loses its culture-specific character and fails to convey a sense of foreignness (Rebechi 2012). Although 'cultural filtering' (Chesterman 1997) could simplify the reader's task by naturalizing cultural terms, it would at the same time frustrate the expectations to taste a foreign flavor.

Mistranslation is also disturbing and abundant. One example will suffice. In a leaflet published by Brazil's Ministry of Culture (Ministério da Cultura 2014), one of the ingredients of *quentão*, a drink served during the June festivals of Brazil and prepared with *cachaça*, sugar, and spices, was translated as 'harpsichord', instead of 'clove'. The confusion is probably due to the polysemy of the word *cravo* in Portuguese, and to the fact that general language dictionaries usually provide a list of decontextualized equivalents, as is the case of the entry for the word *cravo*, translated in the *Webster's Portuguese English Dictionary* (2007) as 'horseshoe nail; spike; corn, callus; comedo (blackhead); spinet, harpsichord, clavichord; carnation, clove-pink' (s.v. *cravo* (m)). As we can see, the dictionary does not provide definitions or examples which could help the researcher to distinguish among the different senses of the equivalents listed, including the musical instrument, mistakenly chosen by the translator. To make things worse, culinary translation is usually relegated to laypersons, since it is traditionally considered an easy task (Teixeira 2004). Although some may believe that recipes can be easily translated, they are now recognized as texts which can reveal important aspects of the source culture, thus demanding more than just terminological expertise and a search for equivalence in order to be appropriately translated. In other words, cooking recipes may demand that the translator be somehow acquainted with the source culture.

Another strategy frequently used in translating Brazilian food terms into English regards the titles of recipes. Based on the traditional narrative framework proposed by Labov (1972), Cotter (1997) compares the title of a recipe to an abstract, which means that it provides the reader with an overview of what follows. This is true for informative or descriptive titles, such as *bolo de coco* 'coconut cake', *arroz com pequi* 'rice with pequi' and *sopa de legumes* 'vegetable soup'. However, Brazilian recipes may often have idiomatic, non-compositional titles, which cannot be translated literally, as this would compromise communication. Literal translations such as *angel's double chin*, *brigadier* and *little peasant girl*, only to mention a few found in the English subcorpus (detailed in

4.1), would certainly create a quite different image in the foreign reader's mind than *papo de anjo*, *brigadeiro* and *caipirinha* do for Brazilians.

As should be clear by now, we believe that identifying and rendering cultural markers appropriately in the target language should be a major concern of translators so as to maintain the specificities of the culture represented in the text. Specialized reference works can help to achieve this goal as they are meant to be a more focused source of terminology retrieval for specific areas, besides contributing with terminology consistency and conventionality.

### 3. BRAZILIAN PORTUGUESE-ENGLISH CULINARY REFERENCE WORKS

Translation studies are frequently divided into two large –and supposedly differing– categories: (i) literary and (ii) technical and scientific. After decades in which the former dominated scholarly attention, the late twentieth and early twenty-first centuries have witnessed an increasing interest in the so-called specialized texts. Actually, a clear-cut distinction between general, literary or specialized texts is questionable. Mayoral Asensio (2007) claims that there is no boundary distinguishing general from specialized language since any act of communication might contain, albeit at different levels, elements of general and specialized languages. *Dona Flor e seus Dois Maridos* (Amado 1966) is a good example of hybridity.<sup>3</sup> Dona Flor, the main character of the novel, is a culinary instructor in Salvador, in the state of Bahia and, hence, traditional Bahian dishes, such as *abará* (puréed black-eyed peas seasoned with salt, onion, dried shrimp and *dendê* oil, wrapped in banana fronds and steamed) and *acarajé* are frequently mentioned in the novel. Nevertheless, such cultural markers have constantly lost their identity in translation as well as in texts originally written in English. We believe that one of the reasons for this problem is directly linked to the reference works available.

We looked into some of the (few) Portuguese-English reference works addressing cookery in order to analyze the extent to which they would help translators and writers. The works we examined were the *Dicionário de Termos Gastronômicos em 6 Idiomas* (Saldanha 2015), *Glossário de Gastronomia: Português-Inglês/Inglês-Português* (Klie 2006), *Vocabulário para Culinária: Inglês-Português* (Teixeira and Tagnin 2008) and

---

<sup>3</sup> See Azenha (1999: 49) for a discussion of hybrid forms.

*Dicionário Gastronômico: Português-Espanhol-Inglês-Alemão-Francês-Italiano* (Carli and Klotz 2007). A summary of the content of the aforementioned works is shown in Table 1 below.

Title	Direction	Equivalent	Phraseology	Definition	Example
<i>Dicionário de Termos Gastronômicos</i>	Portuguese-English/Spanish/French/Italian/German	Yes	No	No	No
<i>Glossário de Gastronomia</i>	Portuguese-English English-Portuguese	Yes	No	No	No
<i>Vocabulário Para Culinária: Inglês-Português</i>	English-Portuguese	Yes	Yes	Yes	Yes
<i>Dicionário Gastronômico</i>	Portuguese-Spanish/English/German/French/Italian	Yes	No	No	No

Table 1: Elements of some Portuguese-English reference works

Except for Teixeira and Tagnin (2008), who drew upon authentic recipes and analyzed them semi-automatically to offer not only equivalents for cooking terms, but also collocations and explanations (for example, distinguishing between similar ingredients), the reference works mentioned would not help translators with the use of terms, since no definitions or examples are offered. In order to help fill this gap, we propose a three-level entry consisting of term/equivalent, appositive explanation and encyclopedic information for our *Portuguese-English Glossary of Brazilian Cooking*.

#### 4. CORPUS LINGUISTICS AND TERM EXTRACTION

Defined by McEnery and Hardie (2012: 1) as an “area which focuses upon a set of procedures, or methods, for studying language,” CL encompasses the compilation and exploration of sets of texts (corpora) collected under well-defined criteria and processed by electronic tools (Bowker and Pearson 2002). A methodology based on CL relies on research in authentic texts, analysis of large amounts of data, automatic retrieval of terms, collocations and recurring combinations (clusters). In addition, it facilitates the search for equivalents and definitions.

Contrary to popular belief that anyone who can cook can translate recipes (see Teixeira 2004), these are highly specialized texts, rich in terminology, which require specialized translator training. As our main purpose is to retrieve terms which are typical of Brazilian cooking, as well as possible equivalents, definitions and authentic examples in English, we have compiled a corpus of Brazilian recipes in Portuguese and in English.

#### 4.1. Cooking recipes

Like any specialized text, recipes contain lexical and syntactic specificities, characteristic terminology –*cup, spoon, dice*– and combinations of words (phraseological units) –*bring to a boil, add gradually, stirring constantly*– which immediately evoke the genre (see Bubel and Spitz 2013). Moreover, the instructions normally conveyed by verbs in the imperative –at least in English and in Portuguese– allow us to define recipes as instructional texts. Due to their highly specialized content, cookbooks were chosen for our study corpus. We compiled a comparable corpus, with recipes originally written in Portuguese and in English, and a parallel corpus, with recipes originally written in Portuguese and their translations into English.

In order to identify what characterizes Brazilian cooking, we relied on eleven cookbooks published in the country as of 1990, in order to privilege what is still part of the population’s eating habits. The cookbooks chosen allegedly comprise national or regional recipes – their titles include the words *Brasil, brasileiro(a)*, or the name of a state or region.<sup>4</sup> All the books feature an introductory text in which the authors discuss different aspects of Brazilian culture and cuisine. We also collected metatexts, in the form of prefaces and introductions, which are frequently included in the cookbooks to explain why the authors decided to write such a work, how the recipes were chosen, what characterizes Brazilian cooking, besides explanations about what is typical of each region. These texts proved to be a rich source of term definition extraction, as will be demonstrated in Section 5.2.

The English counterpart of the comparable corpus, that is, recipes originally written in English, comprises eleven cookbooks published in the United States in the same period, whose titles contain the words *Brazil* or *Brazilian*. Despite the numerous recipes available online, we privileged printed cookbooks as a way to track information considered important to this research, such as authorship, location of publisher, publication date, etc. For the comparable corpus, the cookbooks were required to be originally written in Brazilian Portuguese and in North-American English. This compilation criterion obviously restricted the corpus size, as digitizing the books demanded time. According to the introductory texts, which were manually analyzed, the cookbooks that comprise the study corpus are not addressed to professional cooks. In what concerns the English

---

<sup>4</sup> See Rebechi (2015b) for a complete list of the books which the study corpus comprises.

subcorpus, we observed that the authors provide considerably more information about the reason why they decided to compile a cookbook with Brazilian recipes. In general, they claim to have lived in the country to work or to accompany the spouse.

As six of the eleven cookbooks in Portuguese were translated into English, our research also relies on a parallel corpus. This material was digitized to be automatically processed by *WordSmith Tools* (Scott 2012). Tables 2 and 3 summarize the content of our study corpora.

	<b>Portuguese</b>	<b>English</b>
Introductory texts	194,713	148,136
Recipes	234,704	282,977
<b>Total tokens</b>	<b>429,417</b>	<b>431,113</b>

Table 2: Comparable corpus

	<b>Portuguese</b>	<b>English</b>
Introductory texts	51,806	58,468
Recipes	109,221	107,197
<b>Total tokens</b>	<b>161,027</b>	<b>165,665</b>

Table 3: Parallel corpus

#### 4.2. Terminology and phraseology retrieval

In order to identify Brazilian cultural markers related to cooking, and to propose appropriate equivalents, explanations, and encyclopedic information, complemented by examples of use and phraseological units, we started with a quantitative approach, resorting mainly to the keywords provided by *WordSmith Tools*, which was complemented by a qualitative approach, that is, an analysis of the concordance lines in which these keywords occurred.

Although in general keywords are extracted by comparing a reference corpus of general language with a study corpus, in this study our reference corpus was composed of general cooking recipes in Portuguese, so that the comparison revealed terms which are specific to Brazilian cooking, not general cooking. The reference corpus with approximately one million words (Teixeira 2008) consists of home cooking recipes extracted with an offline browser from websites which do not distinguish their recipes according to their places of origin. As previously stated, the study corpus consists of 22 printed cookbooks which had to be digitalized, and this time-consuming task could not be repeated in the construction of the reference corpus. Therefore, we used a corpus from

a different source, but which is made up of the same genre, that is, cooking recipes.

The Portuguese comparable subcorpus revealed a list of single and compound keywords, that is, words which appear statistically more often in the texts analyzed than in the reference corpus. We then proceeded to retrieve key-keywords, or words which are key in two or more texts or corpora (Scott and Tribble 2006). In order to extract terms related to Brazilian cooking, we only considered keywords which recurred in a minimum of two books. This setting was established as a way of discarding elements which appeared systematically in just one cookbook, which could indicate idiosyncrasy.

From the Portuguese subcorpus with a total of eleven texts, we selected one-word and multi-word terms, which constitute the glossary headwords. Table 4 shows the first 20 terms in decreasing order of keyness. When an English equivalent or translation was identified in the comparable or parallel corpus, it is provided in brackets (detailed information about the identification of equivalents is presented in Section 5.1). Singular and plural occurrences were manually lemmatized.

N	Keyword	Texts	%	Overall Freq.
1	<i>farinha de mandioca</i> ‘manioc flour’	11	100	318
2	<i>mandioca</i> ‘manioc/cassava’	11	100	484
3	<i>charque</i> ‘beef jerky’	10	90.91	111
4	<i>coco(s)</i> ‘coconut(s)’	10	90.91	682 (+49)
5	<i>feijão</i> ‘bean’	10	90.91	283
6	<i>leite de coco</i> ‘coconut milk’	10	90.91	370
7	<i>milho</i> ‘corn’	10	90.91	288
8	<i>arroz</i> ‘rice’	9	81.82	524
9	<i>doce</i> ‘sweet’	9	81.82	191
10	<i>farofa</i>	9	81.82	152
11	<i>goma</i> ‘(manioc) starch’	9	81.82	86
12	<i>pirão</i>	9	81.82	124
13	<i>porco</i> ‘pork’	9	81.82	193
14	<i>arroz branco</i> ‘white rice’	8	72.73	98
15	<i>caranguejo(s)</i> ‘crab(s)’	8	72.73	92 (+16)
16	<i>carne-seca</i> ‘dried beef’	8	72.73	102
17	<i>coco ralado</i> ‘grated coconut’	8	72.73	97
18	<i>coentro</i> ‘cilantro/coriander’	8	72.73	334
19	<i>espigas</i> ‘(corn) cobs’	8	72.73	54
20	<i>jambu</i>	8	72.73	94

Table 4: First 20 one-word and multi-word Portuguese key-keywords

It must be said that, although processing a list of key-keywords greatly facilitated the identification of salient terms or candidates as entries in the glossary, a careful look at

their contexts, mainly enabled by the analysis of concordance lines, determined the final selection. Let us take the key-keyword *codorna* ‘quail’ as an example. This poultry could be an ingredient in many dishes, but the analysis of the concordance lines demonstrated that in the Portuguese subcorpus this keyword is part of the cluster *ovo(s) de codorna* ‘quail egg(s)’ in nine out of its ten occurrences. Therefore, the term which refers to the poultry, rather than the eggs, was discarded. Also, strings of words with incomplete meaning, for example, *de codorna*, *farinha de*, etc. were equally ignored. This way, Brazilian cooking terms were manually selected from the keyword lists in Portuguese to compose the glossary entries. Equivalents, explanatory texts, examples, encyclopedic information, and phraseological units were retrieved from the texts in English. The microstructure of the *Portuguese-English Glossary of Brazilian Cooking* is explained in detail in the next section.

## 5. GLOSSARY ENTRIES: MICROSTRUCTURE

To build our entries, we used *TshwaneLex* (Joffe and de Schryver 2004), a software suite for compiling dictionaries. In addition to the basic three-level categories (equivalent, appositive explanation, and encyclopedic information), we customized the tool so that fields for word class, scientific name (to be used for Brazilian flora and fauna terms), example(s), phraseological unit(s), reference(s), and image(s) were also available. The elements which are judged to be most relevant to the translator and writer of culinary texts are detailed below.

### 5.1. Translation equivalent

In terminology, two terms are considered equivalent when they possess full correspondence of meaning and use within the same area of expertise. However, full correspondence is not a very common phenomenon. Often, a given term in the target language will only partly cover the meaning of the term in the source language (Dubuc 1999). When dealing with an area which is rich in cultural references like cooking, the non-equivalence problem becomes even more evident. However, the translator or writer needs to render concepts from one language to the other as appropriately as possible.

As Newmark (1988: 45) points out, “[t]he central problem of translating has always been whether to translate literally or freely.” As previously discussed, it is common sense

now that translation procedures should take into account the function of the translated text for its own readership. In view of the numerous national food items now available all over the world, it seems that people would look for ethnic food driven by their interest in what is characteristic of a foreign culture. Bearing this in mind, we aim to provide the reader with translation equivalents which maintain the foreignness of culture-bound elements. For the compilation of the *Glossary of Brazilian Cooking*, different equivalence strategies were used, based on Newmark's (1988) translation methods.

#### 5.1.1. Transference

Many Brazilian cooking terms are derived from native Indian and African words. As such, they may sound foreign even to Brazilians, who do not necessarily know their primary meaning. One example is *tapioca*, from Tupi *tipioca*, which means 'clot'. The analysis of this keyword in context showed that it is used mainly to refer to (i) a dish (see Figure 1) and (ii) a type of flour derived from manioc, which is the base for this dish. The analysis of the English subcorpus revealed that, as a dish, the term is usually transferred to the target text; as an ingredient, a number of equivalents were used: *tapioca*, *tapioca flour*, *tapioca starch*, *manioc flour*, *manioc starch*, and *cassava flour*. Due to space constraints, we cannot discuss the adequacy of each translation. For the glossary, we opted for keeping the word in its original form, thus allowing some foreignness to shine through the text, as shown in Figure 1.

**tapioca** (noun) **tapioca** A crepe-like dish prepared with moist manioc starch, eaten plain, with butter or filled with savory or sweet ingredients. //Ex.: *With the heat, the mixture will cohere into a thick pancake. Lower the heat and flip the pancake to cook the other side. [Brasil: Gastronomia, Cultura e Turismo]* A genuinely Brazilian delicacy, tapioca was originally a food of the natives and became very popular in the North and Northeast, where it usually replaces bread during breakfast. Very versatile, the basic recipe (the manioc starch is moistened, sieved and sprinkled onto a hot pan so that the starchy grains fuse into a flatbread) is folded or rolled and, due to its almost neutral flavor, may receive sweet or savory fillings such as grated coconut, curd cheese, shredded beef jerky, banana etc. Besides being fat-, gluten- and sugar-free, tapioca is very nutritious. Nowadays, it is possible to find ready-to-use moist manioc starch for tapioca at supermarkets. See also carne-de-sol, queijo-de-coalho, polvilho. Also known as beiju Compare with tapioca<sup>2</sup>

Figure 1: Entry *tapioca* with borrowed equivalent

### 5.1.2. Literal translation

The list of key-keywords in English revealed highly frequent items, such as *coconut milk*, *shrimp* and *coriander*, which can literally translate as *leite de coco*, *camarão* and *coentro*. Even native items, such as *castanha de caju* and *castanha-do-pará*, have been frequently rendered respectively as *cashew nut* (or simply *cashews*) and *Brazil nut*, words with high keyness in the English subcorpus. Figure 2 shows the entry for *leite de coco*, with the equivalent highlighted.

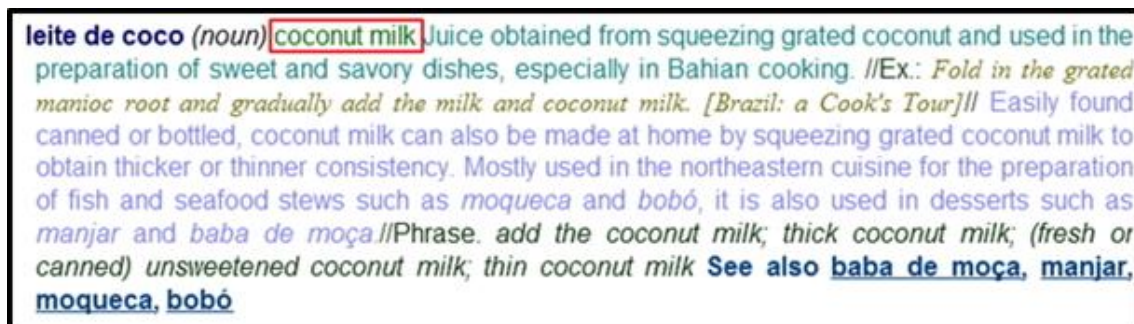


Figure 2: Entry *leite de coco* with literal equivalent

### 5.1.3. Functional equivalent

Even when two languages and cultures do possess similar concepts and, hence, dictionary equivalents, literal translation may not be adequate in all situations. Newmark (1988) explains that choosing a functional equivalent is a common procedure usually applied to cultural words by rendering them with culture-free words, combined or not with transference. A close examination of the keywords in both languages shows that the recipes in North-American English tend to be more technical than the ones in Portuguese. An example is the term *cortador de legumes* ‘vegetable slicer’, frequently rendered in the English subcorpus as *mandoline*, which does have a *prima facie* equivalent in Portuguese, namely *mandolina*. However, this term seems to be restricted to professional cooks. The Portuguese subcorpus has no occurrence of *mandolina*, while the term is described in the book *Chef Profissional* (2011) as an important kitchen utensil.

Considering that the degree of technicity may vary from one language and culture to another, the translator and writer must be very careful about using literal translation to render culture-bound elements. After all, when a term is translated literally, it is often possible to confound the concept referred to with a similar one. Let us take as an example *farinha de milho*, a type of coarse meal made from corn and used basically in the

preparation of two recipes: *cuscuz* and *farofa*. If literally translated as *corn meal*, *cornmeal* or *corn flour*, as identified in the English subcorpus, it might be confused with *fubá*, the main ingredient used in the preparation of polenta, for instance, as if they were interchangeable in recipes. And the reference works available are not helpful in avoiding misunderstandings either.<sup>5</sup>

A search in the parallel corpus revealed the translators' strategy of adding a descriptor as a way of distinguishing these products, which have different characteristics and are used in distinct recipes. Resorting to this procedure, we have added a descriptor, the adjective *flaked*, to account for the coarse texture of this ingredient.<sup>6</sup> Figure 3 highlights the equivalent proposed to render *farinha de milho* in English.

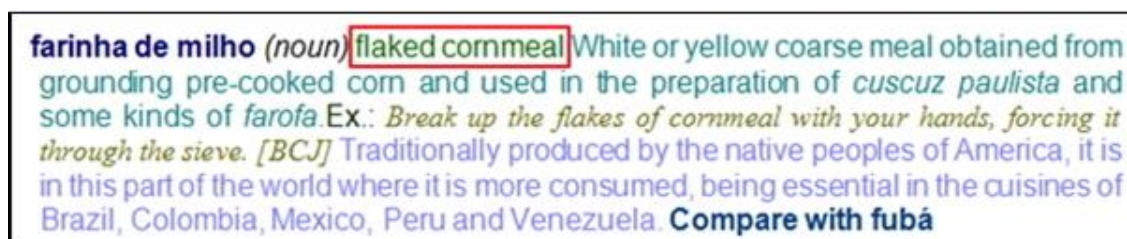


Figure 3: Entry *farinha de milho* with proposed translation equivalent

## 5.2. Appositive explanation

Total equivalence is hardly ever achieved when we deal with cultural markers. Adaptations incur loss of the foreignness which characterizes culture-specific items, whereas transference may result in misunderstandings.

Newmark (1988) explains that descriptive equivalents combine description and function, essential elements in explanation and translation. Here, instead of adopting this concept as a translation procedure, we argue that such a strategy can be combined with transference, literal translation, or functional equivalent to provide translators with a definition which combines both description and function.

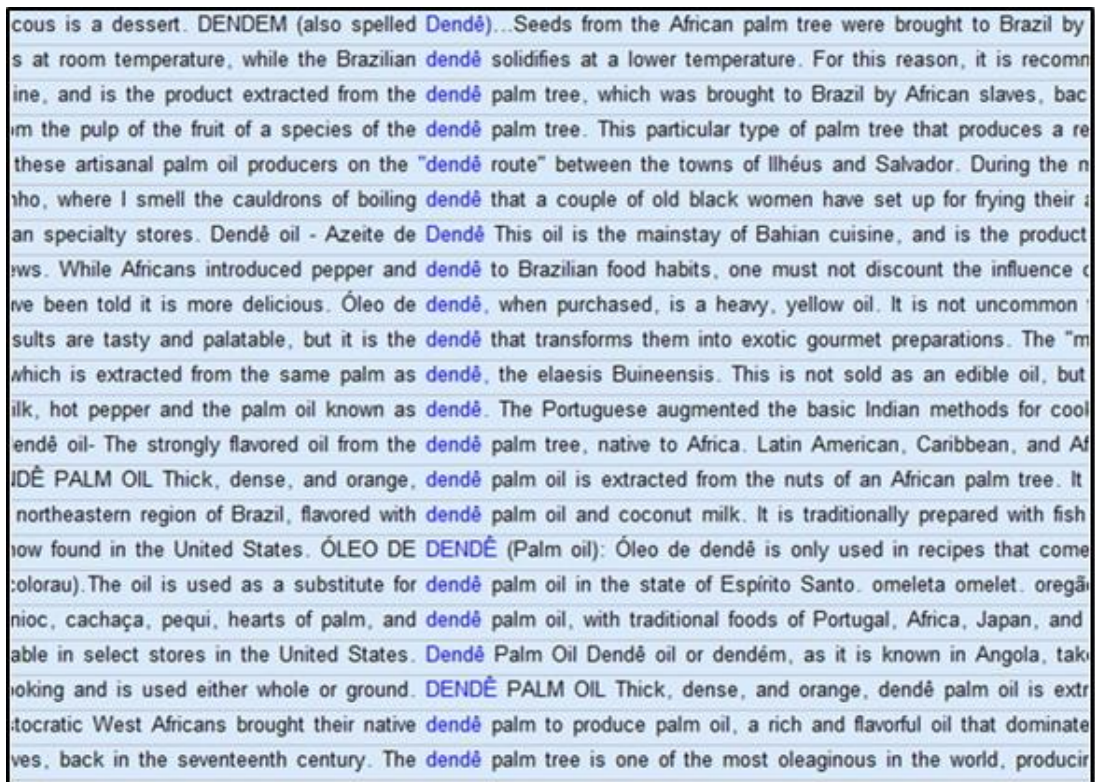
The type of information included in any definition depends essentially on the purpose of the work of reference. Except for Teixeira and Tagnin (2008), the Portuguese-English reference works available do not provide any kind of definition, only translation equivalents. To aid translators and writers, we propose a purpose-specific sentence which

<sup>5</sup> See Rebechi (2015a) for a detailed analysis of how dictionaries and glossaries deal with these ingredients.

<sup>6</sup> See Rebechi (2015a) for a detailed explanation of this translation choice.

allows them to include an elucidation of the cultural term without greatly affecting the fluency of the text. To that end, this appositive explanation consists of a concise phrase, which can either follow the equivalent or be used as a footnote, as a way of providing important information about the term.

The English comparable subcorpus, especially the introductory texts which comprise it, have proved to be excellent sources for the retrieval of relevant information to construct this appositive explanation. Figure 4 shows concordance lines of *dendê* retrieved from these texts.



cous is a dessert. DENDEM (also spelled *Dendê*)...Seeds from the African palm tree were brought to Brazil by  
s at room temperature, while the Brazilian *dendê* solidifies at a lower temperature. For this reason, it is recom  
ine, and is the product extracted from the *dendê* palm tree, which was brought to Brazil by African slaves, bac  
m the pulp of the fruit of a species of the *dendê* palm tree. This particular type of palm tree that produces a re  
these artisanal palm oil producers on the "*dendê* route" between the towns of Ilhéus and Salvador. During the n  
rho, where I smell the cauldrons of boiling *dendê* that a couple of old black women have set up for frying their a  
an specialty stores. Dendê oil - Azeite de *Dendê* This oil is the mainstay of Bahian cuisine, and is the product  
ws. While Africans introduced pepper and *dendê* to Brazilian food habits, one must not discount the influence c  
ve been told it is more delicious. Óleo de *dendê*, when purchased, is a heavy, yellow oil. It is not uncommon  
sults are tasty and palatable, but it is the *dendê* that transforms them into exotic gourmet preparations. The "m  
which is extracted from the same palm as *dendê*, the *elaesis Buineensis*. This is not sold as an edible oil, but  
ilk, hot pepper and the palm oil known as *dendê*. The Portuguese augmented the basic Indian methods for cool  
endê oil- The strongly flavored oil from the *dendê* palm tree, native to Africa. Latin American, Caribbean, and Af  
IDÊ PALM OIL Thick, dense, and orange, *dendê* palm oil is extracted from the nuts of an African palm tree. It  
northeastern region of Brazil, flavored with *dendê* palm oil and coconut milk. It is traditionally prepared with fish  
ow found in the United States. ÓLEO DE *DENDÊ* (Palm oil): Óleo de *dendê* is only used in recipes that come  
colorau).The oil is used as a substitute for *dendê* palm oil in the state of Espírito Santo. omeleta omelet. oregã  
nioc, cachaça, pequi, hearts of palm, and *dendê* palm oil, with traditional foods of Portugal, Africa, Japan, and  
able in select stores in the United States. *Dendê* Palm Oil *Dendê* oil or *dendém*, as it is known in Angola, tak  
oking and is used either whole or ground. *DENDÊ* PALM OIL Thick, dense, and orange, *dendê* palm oil is extr  
stocratic West Africans brought their native *dendê* palm to produce palm oil, a rich and flavorful oil that dominate  
ves, back in the seventeenth century. The *dendê* palm tree is one of the most oleaginous in the world, producir

Figure 4: Concordance lines of *dendê* in introductory texts of cookbooks written in English

As can be seen, the word *dendê* is used to refer to the tree, to the seeds, and to the oil (*azeite de dendê*), the latter being used in cooking. Besides, the titles of the recipes which feature *dendê* also guided us in building the definitions. The concordance lines of *dendê* showed that it is usually present in Bahian dishes, such as *bobó*, *moqueca*, *vatapá* and *acarajé*. All this information can be used to build an appositive explanation (see Figure 5).

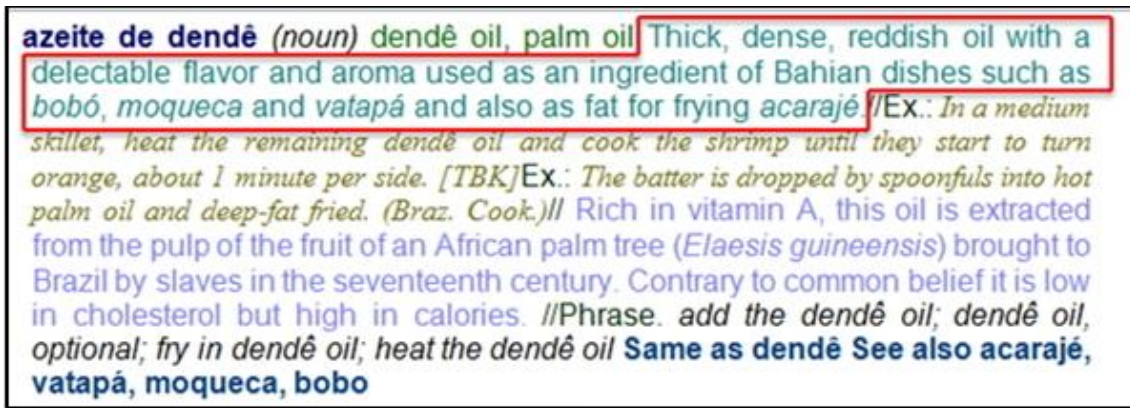


Figure 5: Entry *azeite de dendê* with a highlighted appositive explanation

Because *azeite de dendê* is a typical Brazilian ingredient, its explanation will privilege texture, color, flavor and culinary uses. When a term may be familiar to other cultures the explanation will highlight its use in the Brazilian cuisine. The explanation for *amendoim* ‘peanut’, for example, reads ‘seed used roasted in the preparation of savory Bahian dishes such as *caruru*, *vatapá* and *xinxim*, and also in sweet recipes such as *pé-de-moleque*’.

### 5.3. Example

Again, except for Teixeira and Tagnin (2008), examples are not included in the cooking reference works analyzed. Besides, even when offered (mainly in general language dictionaries), authenticity is not usually sought. However, we believe examples should be authentic, retrieved from naturally occurring texts, thus providing instances of authentic usage of the term, shedding light on the context and even sometimes offering collocations and other phraseological units in which the term may occur. In the *Glossary of Brazilian Cooking* we offer actual examples retrieved from our corpus. The examples are chosen so as to add further information to enhance the reader’s understanding of the term. Figure 6 highlights the examples chosen for the entry *azeite de dendê*. The examples are given credit by using the initials of the books from which they were extracted.<sup>7</sup>

<sup>7</sup> A list with complete references of the books is provided in a separate tab, along with information about the compilation of the glossary.

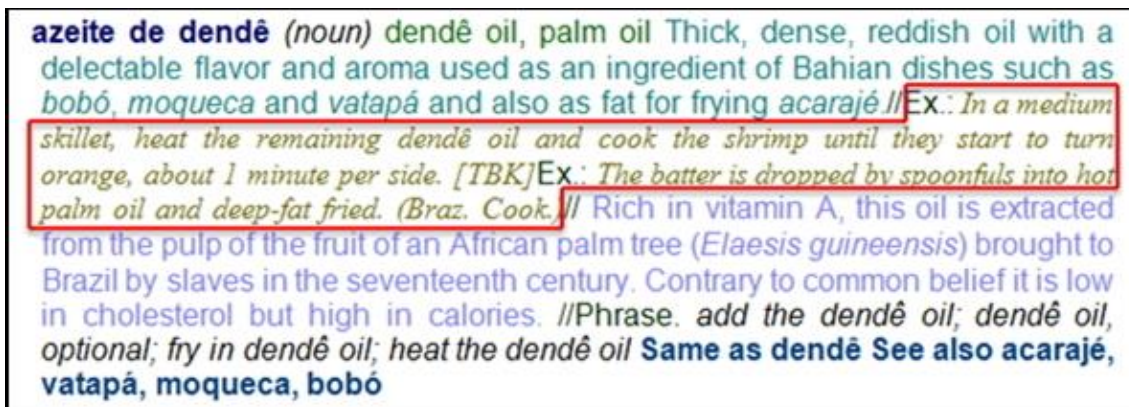


Figure 6: Entry *azeite de dendê* with examples highlighted

#### 5.4. Encyclopedic information

Occasionally the appositive explanation may not be sufficient for a broad understanding of the term in its cooking context. The feature ‘encyclopedic information’ aims at giving the reader extra information about the term, but strictly related to its use in cooking. For example, in the entry *azeite de dendê*, it is important to provide basic nutritional data, such as the fact that the ingredient is rich in vitamin A, and high in calories but not in cholesterol, as many believe. In some cases, we include historical information and links to other reference sources and even to recipes. Figure 7 highlights this element in the entry *azeite de dendê*.

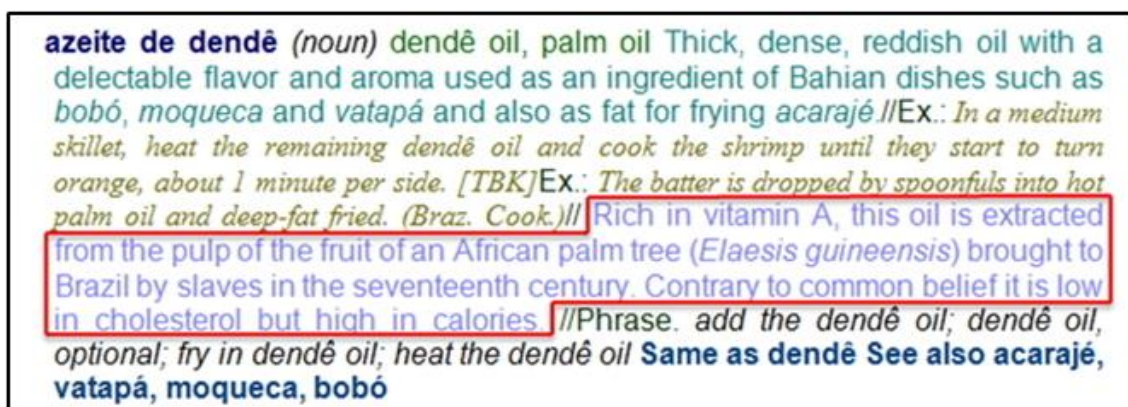


Figure 7: Encyclopedic information for the entry *azeite de dendê*

#### 5.5. Phraseological units

Recent developments in the field of terminology have led to a growing interest in phraseology, since in general translators and specialized writers need to use the terms in

context, not in isolation, to produce fluent texts in the target language. The frequent co-occurrence of words is known by various names: n-grams, multi-word units, clusters, and lexical bundles are some, but not all refer to co-occurrences that convey complete meanings. *WordSmith Tools* generates clusters, which are patterned combinations of words, but not necessarily complete units of meaning. For example, a frequently occurring combination of words such as *devein and* is not understood as complete, but *shell, devein and wash shrimp* is.

Access to specialized phraseological units enables translators to write fluent texts in any given domain, since equivalence is not the only difficulty involved in the translation of cultural markers. Let us take the term *cebola* as an example. A translator would hardly have any difficulty in rendering it as *onion* in the English text. Nevertheless, when faced with the phraseology *cebola cortada em quatro*, the professional, if tempted to use a literal translation strategy, would produce a phrase such as *onion cut in four*, which is not a recurring combination in English. A quantitative analysis will provide us with patterns such as *onion, quartered*; *onion, diced*; *onion, finely chopped*, etc., from which the translator can select the most suitable one (*onion, quartered* in this case).

In addition to the fields mentioned, the *Glossary of Brazilian Cooking* also features cross-references to indicate the semantic-conceptual relations between terminological units. For example, in the entry for *camarão*, a reference such as “Compare with *pitu*” helps to distinguish the term *camarão* ‘shrimp’ with a type of freshwater shrimp, whereas the reference “See also *abará*, *acarajé*, etc.” will refer the reader to dishes in which the term is a key ingredient.

Images may also be helpful in clarifying the meaning of cultural markers. Therefore, this aid has been included whenever deemed relevant. Figure 8 shows a complete entry, using the term *camarão* as an illustration.

**camarão** (*noun*) **shrimp** Sea shellfish sold dried or fresh to be cooked or fried, used as an essential ingredient in many Afro-Bahian dishes such as *vatapá*, *acarajé*, *bobó*, *moqueca*, *abará* etc.//Ex.: Shell, devein and clean shrimp and sauté in hot oil with onions, tomato and green pepper. [Brazilian Cooking] Ex.: Add the shrimp and cook until pink, stirring constantly. [Brazil: a Culinary Journey]// Appreciated all over Brazil, especially by the seashore, shrimp is found in many varieties and sizes and is also used for garnishing. During preparation, stir-fry, cook or sauté shrimp for about two minutes or until they turn pink. Overcooking leads to rubbery and flavorless shrimp.//Phrase. shell, devein and wash shrimp; shrimp, shelled; shrimp, peeled and deveined; (ground) dried shrimp; shrimp filling; shrimp broth Compare with *pitu*, *aviú* See also *abará*, *acarajé*, *caruru*<sup>1</sup>, *vatapá*, *moqueca*



Figure 8: Complete entry of the *Portuguese-English Glossary of Brazilian Cooking* for *camarão*

## 6. CONCLUDING REMARKS

Given that texts about Brazilian cuisine, written in or translated into English, often suffer from misleading or distorting lexical choices when it comes to ethnic food terms, and assuming these issues could be minimized if professionals had access to more comprehensive and reliable terminological resources, this paper proposes the compilation of a *Portuguese-English Glossary of Brazilian Cooking* aimed at translators and specialized writers. To accomplish our aims, we have relied on a CL approach.

In addition to appropriate equivalents, we propose a few features aimed at facilitating the translation of cultural terms. The main one, which represents an innovation in Brazilian culinary reference materials, is an appositive explanation, a short text which can be inserted in the translation without affecting its fluency or, if the writer so chooses, used as a footnote. Next, authentic examples taken from our corpora are provided. In addition, we offer extra information for the term, in case the translator or reader is interested in learning more about it. Therefore, we propose the inclusion of encyclopedic information, which may give historical details of the ingredient or dish, add a recipe or links to other reference sources. Phraseology is also addressed showing the appropriate use of a term in relation to its most frequently occurring collocates so as to add fluency to the text.

We believe that a carefully compiled reference source with cultural information would help not only to prevent mistakes but also to recover cultural markers in the target text, whether predominantly literary or specialized. Besides, we hope that our proposal for a three-level entry in a specialized glossary will find application in other terminological areas, especially those focused on the compilation of reference works which address cultural items.

#### REFERENCES

- Amado, Jorge. 1966. *Dona Flor e seus Dois Maridos*. São Paulo: Companhia das Letras.
- Azenha Jr., João. 1999. *Tradução Técnica e Condicionantes Culturais: Primeiros Passos para um Estudo Integrado*. São Paulo: Humanitas.
- Barnes, Julian. 2003. *The Pedant in the Kitchen*. London: Atlantic Books.
- Bowker, Lynne and Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpus*. London: Routledge.
- Brien, Donna L. 2007. Writing about food: Significance, opportunities and professional identities. In Jen Webb and Jordan Williams eds. *Proceedings of the 12<sup>th</sup> Annual Conference of the Australian Association of Writing Programs*. Canberra: University of Canberra, 1–15.
- Bubel, Claudia and Alice Spitz. 2013. The way to intercultural learning is through the stomach. In Cornelia Gerhardt, Maximiliane Frobenius and Susanne Ley eds. *Culinary Linguistics: The Chef's Special*. Amsterdam: John Benjamins, 157–187.
- Capatti, Alberto and Massimo Montanari. 1999. *La Cucina Italiana: Storia di una Cultura*. Roma-Bari: Editori Laterza.
- Carli, Francesco and Eliane Klotz eds. 2007. *Dicionário Gastronômico: Português-Espanhol-Inglês-Alemão-Francês-Italiano*. São Paulo: Contorno.
- Chesterman, Andrew. 1997. *Memes of Translation: The Spread of Ideas in Translation Theory*. Amsterdam: John Benjamins.
- Cotter, Colleen. 1997. Claiming a piece of the pie: How the language of recipes defines community. In Anne L. Bower ed. *Recipes for Reading: Community Cookbooks, Stories, Histories*. Massachusetts: University of Massachusetts Press, 51–72.
- Dubuc, Robert. 1999. *Manual Prático de Terminología* (translated by Ileana Cabrera). Providencia: RiL.
- Gerhardt, Cornelia. 2013. Food and language –language and food. In Cornelia Gerhardt, Maximiliane Frobenius and Susanne Ley eds. *Culinary Linguistics: The Chef's Special*. Amsterdam: John Benjamins, 3–49.
- Gerhardt, Cornelia, Maximiliane Frobenius and Susanne Ley eds. 2013. *Culinary Linguistics: The Chef's Special*. Amsterdam: John Benjamins.
- Instituto Americano de Culinária. 2011. *Chef Profissional* (fourth edition). São Paulo: Senac.
- Joffe, David and Gilles-Maurice de Schryver. 2004. TshwaneLex, a state-of-the-art dictionary compilation program. In Geoffrey Williams and Sandra Vessier eds. *Proceedings of the Eleventh EURALEX International Congress*. Lorient Cedex: Université de Bretagne Sud, 99–104.
- Jurafsky, Dan. 2014. *The Language of Food: A Linguist Reads the Menu*. New York: W. W. Norton & Company.

- Klie, Virginia. 2006. *Glossário de Gastronomia: Português-Inglês / Inglês-Português*. Rio de Janeiro: Disal.
- Labov, William. 1972. *Language in the Inner City: Studies in the Black English Vernacular*. Philadelphia: University of Pennsylvania Press.
- Mayoral Asensio, Roberto M. 2007. Specialised translation: A concept in need of revision. *Babel: Revue Internationale de la Traduction* 53: 48–55.
- McEnery, Tony and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
- Ministério da Cultura. 2014. *Aromas, Cores e Sabores do Brasil*. [http://www.copa2014.gov.br/sites/default/files/livreto\\_web17062013.pdf](http://www.copa2014.gov.br/sites/default/files/livreto_web17062013.pdf) (1 April, 2020.)
- Newmark, Peter. 1988. *A Textbook of Translation*. New York: Prentice-Hall.
- Nord, Christiane. 2001. *Translating as a Purposeful Activity: Functionalist Approaches Explained*. Manchester: St Jerome.
- Nord, Christiane. 2012. Functional approaches to translation. In Carol A. Chapelle ed. *The Encyclopedia of Applied Linguistics*. Hoboken: Blackwell Publishing, 2223–2228.
- Pearson, Jennifer. 1998. *Terms in Context*. Amsterdam: John Benjamins.
- Rebechi, Rozane R. 2012. ‘Cachaça’ na tradução de obras literárias brasileiras para a língua inglesa. *Tradterm* 20: 95–110.
- Rebechi, Rozane R. 2015a. Tracing English equivalents of Brazilian Portuguese cooking vocabulary: A corpus-based study. In Sattar Izwaini ed. *Papers in Translation Studies*. Newcastle upon Tyne: Cambridge Scholars, 154–178.
- Rebechi, Rozane R. 2015b. *A Tradução da Culinária Típica Brasileira para o Inglês: Um Estudo sob o Enfoque da Linguística de Corpus*. São Paulo, SP: The University of São Paulo dissertation.
- Saldanha, Roberta M. 2015. *Dicionário de Termos Gastronômicos em 6 Idiomas*. Rio de Janeiro: Senac.
- Scott, Mike. 2012. *Wordsmith Tools Version 6.0*. Stroud: Lexical Analysis Software.
- Scott, Mike and Christopher Tribble. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.
- Teixeira, Elisa D. 2004. *Receita Qualquer um Traduz. Será? A Culinária como Área Técnica de Tradução*. São Paulo, SP: The University of São Paulo dissertation.
- Teixeira, Elisa D. 2008. *A Linguística de Corpus a Serviço do Tradutor: Proposta de um Dicionário de Culinária Voltado para a Produção Textual*. São Paulo: The University of São Paulo dissertation.
- Teixeira, Elisa D., and Stella E. O. Tagnin. 2008. *Vocabulário para Culinária: Inglês-Português*. São Paulo: SBS.
- Temmerman, Rita and Danièle Dubois eds. 2017. Food and terminology: Expressing sensory experience in several languages. Special issue of *Terminology* 23/1.
- Tigner, Amy L. and Allison Carruth. 2018. *Literature and Food Studies*. London: Routledge.
- Webster's Portuguese-English Dictionary* (eighteenth edition). 2007. Rio de Janeiro: Record.
- Zavaglia, Adriana, João Azenha and Tinka Reichmann. 2011. Cultural markers in LSP translation. In Klaus-Dieter Baumann ed. *Fach - Translat - Kultur: Interdisziplinäre Aspekte der Vernetzten Vielfalt*. Berlin: Frank und Timme, 785–808.

*Corresponding author*

Rozane Rebechi

Avenida Bento

Gonçalves, 9500,

Porto Alegre, RS, Brazil

e-mail: rozane.rebechi@ufrgs.br

received: January 2020

accepted: April 2020

# EusTimeML: A mark-up language for temporal information in Basque

Begoña Altuna - María Jesús Aranzabe - Arantza Díaz de Ilarraza  
University of the Basque Country / Spain

**Abstract** – We present EusTimeML, a mark-up language for temporal information in texts written in Basque. It is compliant with the TimeML specifications, while offering some adapted attributes and attribute values in order to represent the language-specific features of Basque. In particular, alterations have been carried out for verb tense, aspect and modality coding, as well as for time expression and signal annotation. EusTimeML also provides a major extension to the existing TimeML schemes, since the attributes and values for factuality annotation have been added to the existing temporal information annotation scheme. EusTimeML has been used to annotate the *EusTimeBank Corpus*, the news and history narratives corpus that has been used as the gold standard in temporal information processing in Basque.

**Keywords** – temporal information processing; Basque; mark-up language; annotation; TimeML

## 1. INTRODUCTION

Natural Language Processing (NLP) aims at getting the deepest textual understanding, for which, after mastering morphosyntactic analysis, the focus has been put on semantic and discourse information. Temporal information is an integral part of those areas as it conveys the information of what is narrated in text while providing information to arrange narratives along a temporal axis. This information is of utmost relevance to the development of automatic systems that benefit from knowing the chronological ordering of events in texts, such as chronology creation (Bauer *et al.* 2015), event prediction (Radinsky and Horvitz 2013) and event forecasting systems (Kawai *et al.* 2010), among others.

Specifically, temporal information conveys the information of what happens (events narrated) and the times in which they happen (time expressions), as well as the temporal relations (simultaneity, precedence, etc.) between them. For example, in the sentence in (1), one can learn that there was a toilet paper theft (event) last month (time expression) after (temporal relation) there were shortages (event).

- (1) Last month, armed robbers stole pallets of toilet paper in Hong Kong following panic-buying induced shortages.

That temporal information is collected in corpora that are annotated following structured formats, e.g., the eXtended Mark-up Language (XML), which make the information in the texts machine-readable. Mark-up languages provide a set of tags to classify the different elements in the text, as well as a set of attributes to describe the relevant linguistic features of those elements.

For the annotation of temporal information in Basque, we have created EusTimeML, a TimeML-compliant mark-up scheme (Pustejovsky *et al.* 2003a). It provides tags for events, time expressions and the relations that hold between them in XML format. As Figure 1 shows, some text strings have been assigned a tag (in green) since those are the elements in text that express temporal information. Additionally, a set of attributes (in purple) represents the main information (in pink) those strings convey.

```
<?xml version="1.0" encoding="UTF-8"?>
<TimeML>
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="http://www.cognitionis.com/corpus/es/tml.xsd"
  <DOCID>10026-First_A380_enters_commercial_service.txt</DOCID>
  <DCT>
    <TIMEX3 tid="t0" type="DATE" value="2007-10-17" temporalFunction="false"
      functionInDocument="CREATION_TIME">2007-10-17</TIMEX3>
  </DCT>
  <TEXT>
    Lehen A380a
    <EVENT eid="e1" class="ASPECTUAL">hasi</EVENT>
    da
    <EVENT eid="e2" class="OCCURRENCE">zerbitzu</EVENT>
    komertziala
    <EVENT eid="e3" class="OCCURRENCE">ematen</EVENT>
    .
    <TIMEX3 tid="t4" type="DATE" value="2007-18-17">2007ko urriaren 17a</TIMEX3>
    . Hegazkingintzaren historian , sekulako
    <EVENT eid="e4" class="OCCURRENCE">lorpena</EVENT>
    <EVENT eid="e5" class="OCCURRENCE">izan</EVENT>
    da : lehen Airbus A380a
    <TIMEX3 tid="t15" type="TIME" value="2007-10-17T18:40">18:40an ( GMT + 8 )</TIMEX3>
    <EVENT eid="e6" class="OCCURRENCE">lurreratu</EVENT>
    da Singapur-en , Changi nazioarteko aireportuan , Airbus-en bidalketa-zentrotik
    <EVENT eid="e7" class="OCCURRENCE">atera</EVENT>
    eta
    <TIMEX3 tid="t17" type="DURATION" value="PT12H">12 orduko</TIMEX3>
    <EVENT eid="e8" class="OCCURRENCE">hegaldia</EVENT>
    <EVENT eid="e9" class="OCCURRENCE">egin</EVENT>
    <SIGNAL sid="s1">ostean</SIGNAL>
    . Hegazkinari 400 bat gonbidatuk
    <EVENT eid="e10" class="OCCURRENCE">egin</EVENT>
    zioten
    <EVENT eid="e11" class="OCCURRENCE">ongietorria</EVENT>
    laster
    <EVENT eid="e12" class="OCCURRENCE">zabalduko</EVENT>
    den Changi Nazioarteko aireportuko 3. terminalean .
  </TEXT>
</TimeML>
```

Figure 1: A text annotated following the EusTimeML mark-up language (simplified annotation)

The text in Figure 1 is part of the *EusTimeBank Corpus* (Altuna *et al.* under revision a) which, in turn, has been used to train and evaluate temporal information processing

tools. The Basque language has a long tradition of linguistic analysis and automatic processing (Alegria and Sarasola 2017) and integrating temporal information processing in the Basque processing pipeline (Otegi *et al.* 2016) has been the major motivation of this work.

This paper is structured as follows. We revisit the most relevant work on temporal information mark-up languages in Section 2. In Section 3, we present the basic features of TimeML, and in Section 4 we describe the most relevant linguistic features of Basque and the adaptations of TimeML that we have instituted to accommodate those features. We discuss the strengths and weaknesses of EusTimeML in Section 5, and we conclude our work in Section 6.

## 2. BACKGROUND

Temporal information processing has attracted the interest of NLP scholars over the last two decades and has experienced a substantial boost since the creation of TimeML (Pustejovsky *et al.* 2003a). In fact, ever since the creation of TimeML, resource generation efforts and system evaluation competitions have multiplied. TimeML has been adapted to multiple languages, tasks and domains, and corpora annotated with TimeML schemes have increased in number.

The first temporal information mark-up languages (Mani and Wilson 2000; Ferro *et al.* 2003) only dealt with time expressions, for which the TIMEX and TIMEX2 tags respectively were created. These two tags also offered a set of basic attributes to code the main information expressed by time expressions, such as the normalised value and the granularity of the time expression. TimeML (Pustejovsky *et al.* 2003a), instead, made a qualitative leap in temporal information annotation, as this mark-up language offered tags for all the elements taking part in the expression of temporal information (see Section 3).

TimeML is now an ISO<sup>1</sup> standard (Pustejovsky *et al.* 2010) used in the annotation of many temporally annotated corpora such as *TimeBank* (Pustejovsky *et al.* 2003b, 2006), the *THYME Corpus* (Styler *et al.* 2014), the *PHEME Tweet Corpus* (Derczynski and Bontcheva 2014) and the *Event StoryLine Corpus* (Caselli and Vossen 2017), among others. Moreover, TimeML has been adapted to address some special annotation

---

<sup>1</sup> International Standards Organisation.

needs. TimeML-strict (Derczynski *et al.* 2013) aims at reducing annotation ambiguity and TimeML-Dense (Cassidy *et al.* 2014) offers the opportunity to create denser temporal relation graphs, while Mostafazadeh *et al.* (2016) use a reduced version of TimeML to annotate the *ROCStories Corpus*.

TimeML has also been developed for many languages, as it is considered a *de facto* standard. Among other languages, TimeML schemes are available for French (Bittar 2010), Italian (Caselli *et al.* 2011), Portuguese (Costa and Branco 2012) Romanian (Forăscu and Tufiş 2012), Spanish (Saurí *et al.* 2009, 2010; Saurí 2010) and Catalan (Saurí and Pustejovsky 2009, 2010; Saurí 2010) and Korean (Jeong *et al.* 2015).

Nonetheless, TimeML is not the only mark-up language that has been developed to address temporal information. TEMANTEX (Wonsever *et al.* 2015) merges event annotation and factuality annotation. In the mark-up language developed for the *NewsReader project* (Minard *et al.* 2016), in turn, temporal information is tagged as in TimeML, but causality relations and entity co-reference are also considered. PLIMEX (Kocoń and Marcińczuk 2015) addresses time expressions in Polish and follows TimeML guidelines quite narrowly. Finally, Ning *et al.* (2018) created a mark-up language that focuses on the extraction of relevant information for timeline construction. This mark-up language complies with most of the TimeML decisions, while it offers a much richer annotation for intrasentential temporal relations.

### 3. TIMEML

The TimeML mark-up language was specifically created to annotate events, time expressions and the temporal relations between them in text (Pustejovsky *et al.* 2010). For that, the following set of tags was defined, one for each element concerning temporal information or type of relation:

- `<EVENT>` for events: actions and situations that happen or occur, as in (2).<sup>2</sup>
  - (2) Numerous conspiracies have `<EVENT>appeared</EVENT>` since the `<EVENT>outbreak</EVENT>`.
- `<TIMEX3>` for temporal expressions that convey date, time, duration or set information, as in (3).

---

<sup>2</sup> TimeML foresees single-token annotations for events.

- (3) Cases of the new coronavirus emerged in Wuhan <TIMEX3>late last year</TIMEX3>.
- <SIGNAL> for sections of text, most commonly function words, that indicate the type of relation among temporal objects, as in (4).
- (4) Numerous conspiracies have appeared <SIGNAL>since</SIGNAL> the outbreak.
- <TLINK> for temporal relations between two events, two time expressions or an event (in bold) and a time expression (in italics), as in (5).
- (5) Numerous conspiracies (ei1) have **appeared** (ei2) since the *outbreak* (ei3).  
 <TLINK eventInstanceID="ei2" relatedToEvent="ei3" relType="BEGUN\_BY"/>
- <ALINK> for aspectual relations between an aspectual event (in bold) and its subordinated event (in italics), as in (6).
- (6) Several patent documents **started** (ei1) to *circulate* (ei2) on Twitter. <ALINK eventInstanceID="ei1" relatedToEvent="ei2" relType="START"/>
- <SLINK> for subordination relations between a main event (in bold) and its subordinated event (in italics), as in (7).
- (7) Ms Mengyun apologised (ei1), **saying** (ei2) she was "just *trying* (ei3) to introduce (ei4) the life of local people". <SLINK eventInstanceID="ei2" relatedToEvent="ei3" relType="EVIDENTIAL"/>

These tags contained a set of attributes that coded or normalised the temporal information conveying features of the temporal objects and relations. Table 1 presents the attributes in TimeML for event features. In this case, four types of attributes can be identified according to the type of information they represent: i) event ID (*eid*) and event instance ID (*eiid*) offer identification information; ii) class (*class*) offers event classification information; iii) tense (*tense*) and aspect (*aspect*) offer temporal information; and, finally iv) part-of-speech (*pos*), polarity (*polarity*) and modality (*modality*) offer other relevant linguistic information. Those attributes may get different types of values. Some attribute values can be strings (CDATA) or integers, such as in *eid*, while others get their values from a list of pre-established options, such as for *class*, *tense* and *aspect*.

Attributes and attribute values for the remaining tags have also been defined in TimeML. In the case of TIMEX3 the most relevant tags express the type and normalised value of the time expressions, and SIGNAL tags do not get any attributes. For the

relations, the source, target and relation type are specified. The complete description of the TimeML tags, attributes and attribute values can be found in TimeML Working Group (2010).

Event attributes	Values
Event ID ( <i>eid</i> )	e<integer>
Event instance ID ( <i>eiid</i> )	ei<integer>
Class ( <i>class</i> )	REPORTING, PERCEPTION, ASPECTUAL, I_ACTION, I_STATE, OCCURRENCE
Tense ( <i>tense</i> )	PAST, PRESENT, FUTURE, NONE, INFINITIVE, PRESPART, PASTPART
Aspect ( <i>aspect</i> )	PROGRESSIVE, PERFECTIVE, PERFECTIVE_PROGRESSIVE, NONE
Part of speech ( <i>pos</i> )	ADJECTIVE, NOUN, VERB, PREP, OTHER
Polarity ( <i>polarity</i> )	NEG, POS
Modality ( <i>modality</i> )	CDATA

Table 1: Attributes of the TimeML <EVENT> tag and their possible values

#### 4. LANGUAGE-SPECIFIC ISSUES AND EXTENSIONS TO TIMEML

EusTimeML follows most of the standards in TimeML, namely, it preserves the token-level annotation system and all the tags proposed in TimeML, as well as the attributes and values that code temporal information. Additionally, it also shares almost all the attributes for linguistic features and their values. Nonetheless, Basque has some language-specific issues (see Section 4.1) that have conditioned the adaptation of TimeML to Basque (see Section 4.2), for which a series of attribute values has been altered. Furthermore, EusTimeML offers a set of attributes to address some supplementary information so as to increase the amount and variety of information it encodes (see Section 4.3). These all have contributed to the final version of EusTimeML (see Section 4.4).

##### 4.1. Language-specific issues

Basque is a non-Indo-European language isolate, and thus it does not share many of the linguistic features of its neighbouring languages. In particular, many of its morphosyntactic features differ from the features in neighbouring languages and, hence, specific research for processing Basque is usually needed, as choices made for other languages cannot be applied straightforwardly.

For example, Basque is a highly agglutinative language in which information commonly expressed by prepositions in neighbouring languages is expressed by a rich set of postpositions attached to lemmas, as can be seen in the sentence in (8). This feature is extremely relevant in temporal information processing as lemmas accompanied by spatio-temporal declension cases are very frequent in temporal information expressions.

(8) *Sorosleek iluntzetik egunsentira etengo dituzte erreskate-operazioak.*

Rescuers sunset.ABL sunrise.ALL stop.FUT aux.PRES rescue.operations.

‘Rescuers will stop rescue operations from sunset to sunrise.’

In (8) there are two time expressions: *iluntzetik* ‘from sunset’ and *egunsentira* ‘to sunrise’. *Iluntze* and *egunsenti* mean ‘sunset’ and ‘sunrise’, respectively, while the suffix *-etik* expresses the ablative case and *-ra* represents the allative case.

Verbal conjugation also represents a major difference between Basque and other languages. In Basque, there is a short list of single-word verb forms, typically to express punctual aspect, whereas most of the tensed verb forms are periphrastic. The lexical meaning of the verb and aspect are expressed in the main verb, while the auxiliary verb expresses tense and agreement with the persons taking part in the event, as well as mood and modality.

Looking at the sentence in (8) again, one may notice that the verb *etengo dituzte* (‘will stop’) also shows the rich morphology of Basque. The suffix *-go* expresses the future aspect of the verb and the auxiliary *dituzte* represents the present time tense (*d-*) as well as the concordance with the object (*erreskate-operazioak.3PL*), *-it-* and *-z-*, and the subject (*sorosleek.3PL*), *-te*.

As just mentioned, the future meaning of a verb is considered an aspectual value in Basque, whereas in many European languages future events are expressed by the future tense. This makes it possible to understand the Basque verbal tense as a bi-dimensional present-past feature (Table 2), and verbal aspect as a perfect-future feature (Table 3).

<b>Present</b>		
	Present	Non-present
<b>Past</b>	Non-past Eten dituzte ('They have stopped')	Eten (izan) balituzte ('If they had stopped')
	Past	Eten zitutzen ('They stopped')

Table 2: Representation of verbal tenses in Basque

<b>Perfectiveness</b>		
	Perfect	Non perfect
<b>Futurity</b>	Non-future Eten dituzte ('They have stopped')	Eteten dituzte ('They stop')
	Future	Etengo dituzte ('They will stop')

Table 3: Representation of verbal aspect in Basque

## 4.2. Adaptations to TimeML

Although the TimeML mark-up language is considered to be a standard for temporal information annotation, each version contains subtle variations to address language or task-specific issues. In EusTimeML, some of the attribute values have been modified to accommodate the analysis of Basque grammar.

### 4.2.1. Time expression and signal annotation

As introduced in Section 4.1, time expressions often get spatio-temporal postpositions and both elements commonly appear as a single token. Those elements are given separated tags in TimeML-styled schemes: one for the time expression (<TIMEX3>) and one for the function word (normally a preposition) expressing a temporal relation (<SIGNAL>). In the case of Basque, instead, as EusTimeML respects the token-level annotation, we decided to annotate the whole word as a time expression, since we believe that the postpositions' relational information can always be recovered from the morphosyntactic parsing.

Nevertheless, free postpositions are also possible in Basque and, in those cases, we decided to assign them a signal tag, as the tags for free postpositions do not interfere with any other tags present in a text. As a consequence, the time expression and signal information annotation according to EusTimeML is represented as in examples (9–10). A similar decision was made for the annotation of events and signals. More precisely, as

only main verbs are given the event tag, the auxiliaries of the periphrastic forms may get signal tags when they contain a temporal postposition, since there is no overlapping tag as in (11).

- (9) Sorosleek <TIMEX3>iluntzetik</TIMEX3>  
 <TIMEX3>egunsentira</TIMEX3> etengo dituzte erreskate-operazioak.  
 ‘Rescuers will stop rescue operations from <TIMEX3>sunset</TIMEX3> to  
 <TIMEX3>sunrise</TIMEX3>’

- (10) Krimean gotortu eta <TIMEX3>1920ko udazkenara</TIMEX3>  
 <SIGNAL>arte</SIGNAL> eutsi zuten.  
 ‘[They] hid in Crimea and the endured <SIGNAL>until</SIGNAL>  
 <TIMEX3>Autumn 1920</TIMEX3>’

- (11) Gerra Zibila Armada Zuria <EVENT>menderatu</EVENT>  
 <SIGNAL>zutenean</SIGNAL> amaitu zen.  
 ‘Civil War ended <SIGNAL>when</SIGNAL> [they]  
 <EVENT>overruled</EVENT> the White Army’

As can be seen in (10), the free postposition *arte* has been assigned a SIGNAL tag. The ablative *-tik* and the allative *-ra* in (9), and the allative *-era* in (10), instead, are part of the TIMEX3 tag to which they are attached. Nevertheless, it should be noted that, in (11), the auxiliary *zutenean* contains the locative suffix *-ean*, but as there is no conflict with any other tags, the token has been assigned a SIGNAL tag according to EusTimeML.

#### 4.2.2. Aspect and tense annotation

The fact that the future is represented by aspect in Basque has led us to define an *ad hoc* set of values for aspect and tense. As in other TimeML-styled schemes, verbal aspect is expressed by the *aspect* attribute and verb tense is represented through the *tense* attribute. The values each attribute can be assigned to and the context have been summarised in Table 4.

TENSE		ASPECT	
Values	Usages	Values	Usages
PRESENT	Events expressed by verbs in the present tense	PERFECT	Events expressed by verbs with perfective aspect
PAST	Events expressed by verbs in the past tense	-PERFECT (NON-PERFECT)	Events expressed by verbs with imperfective aspect
HYPOTHETICAL	Events expressed by verbs in the hypothetical (non-present, non-past) tense	FUTURE	Events expressed by verbs with future aspect
NONE	Events expressed by untensed verbs and non-verbal forms	NONE	Events expressed by verbs with no aspect mark and non-verbal forms

Table 4: Values and usages of the *aspect* and *tense* attributes in EusTimeML

As a consequence, for the sentence in (8) the event *etengo dituzte* (‘will stop’) would be assigned the *aspect* and *tense* values as illustrated in (12), since this is a future verb form. *Erreskate-operazioak* (‘rescue operations’), instead, will be assigned NONE as the value for *aspect* and *tense*, as it is expressed by a noun phrase and the form has no aspect or tense marks.

(12) Sorosleek iluntzetik egunsentira <EVENT aspect="FUTURE"  
tense="PRESENT">etengo</EVENT> dituzte <EVENT aspect="NONE"  
tense="NONE">erreskate-operazioak</EVENT>.

‘Rescuers will <EVENT>stop</EVENT> rescue <EVENT>operations</EVENT>  
from sunset to sunrise’

#### 4.2.3. Modality annotation

The annotation of modality information has been tackled in various ways in the different TimeML-styled mark-up languages. While the Spanish and Catalan TimeML schemes (Saurí and Pustejovsky 2009) do not contain any means for modality annotation, the rest of the analysed mark-up languages do offer the option to express modality through the *mod* attribute. However, while a set of definite values has been defined for the French TimeML (Bittar 2010), in the rest of the analysed mark-up languages the forms found in text are used as values.

In the case of EusTimeML, we have followed the Basque grammar tradition, in which the modal verb *ahal izan/ezin izan* (possibility) and the semi-modal verbs *behar izan* (need or obligation) and *nahi izan* (desire) are considered. Taking this into account, the AHAL, BEHAR and NAHI values have been created for the modality attribute, and we have also used the NONE value for the cases in which no modality is expressed. The

NONE value is the one assigned to the events in (13), as they do not convey any modality information.

(13) Sorosleek                      iluntzetik                      egunsentira                      <EVENT  
       modality="NONE">etengo</EVENT>                      dituzte                      <EVENT  
       modality="NONE">erreskate-operazioak</EVENT>.  
       ‘Rescuers will <EVENT>stop</EVENT> rescue <EVENT>operations</EVENT>  
       from sunset to sunrise’

#### 4.3. Extensions to TimeML: Factuality

The main difference between EusTimeML and other TimeML-styled mark-up schemes relies on the factuality annotation added to EusTimeML (Altuna *et al.* 2018a). Factuality annotation has been closely related to TimeML in works such as Saurí (2008), but EusTimeML is the first TimeML-styled scheme that integrates it. For example, verb aspect and tense, the time expressions related to the events condition, the factuality values of the events, and some subordination relations (evidential, factive or counterfactive, among others) may evidence the factuality value of the subordinated event.

As our final goal is building timelines, factuality information will help us discern between events that effectively do occur and that should, as a consequence, appear on a timeline, events that have not happened, and events that may happen in the future. For this reason, we have opted for a factuality scheme in which we classify events as facts, counterfactuals, or non-factual events when possible.

In EusTimeML, factuality information is coded through a set of event attributes. These attributes are *polarity* (defined also in TimeML), *certainty*, *factuality* itself, and *specialCases*. These attributes and their values are illustrated in Table 5.

Polarity	Certainty	Factuality	specialCases
<i>Grammatical polarity (affirmation or negation) expression</i>	<i>Commitment of the source with the information expressed</i>	<i>Information on whether events correspond to a fact in the world, a possibility or a situation that does not hold</i>	<i>Marking of conditionals and generic statements</i>
POS NEG	CERTAIN UNCERTAIN UNDERSPECIFIED	FACTUAL COUNTERFACTUAL NON-FACTUAL NO FACTUALITY VALUE UNDERSPECIFIED	CONDITIONAL_CONDITION CONDITIONAL_MAIN (main clause) GENERIC NONE

Table 5: Factuality specific attributes and values in EusTimeML

We have represented the factuality information of the events in (8) as shown in example (14).

- (14) Sorosleek iluntzetik egunsentira <EVENT polarity="POS" certainty="CERTAIN" factuality="NON\_FACTUAL" specialCases="NONE">etengo</EVENT> dituzte <EVENT polarity="POS" certainty="CERTAIN" factuality="FACTUAL" specialCases="NONE">erreskate-operazioak</EVENT>.  
 ‘Rescuers will <EVENT>stop</EVENT> rescue <EVENT>operations</EVENT> from sunset to sunrise’

#### 4.4. Final EusTimeML definition and usage

Taking into account the decisions we made, the attributes and values for event annotation in EusTimeML are presented in Table 6. The remaining tags preserve the original TimeML attributes and the only differences in annotation are the ones presented in Section 4.2.1. As a consequence, annotations following EusTimeML remain easily transferable and comparable to other annotations carried out following any of the TimeML-styled schemes.

Event attributes	Values
Event ID ( <i>eid</i> )	e<integer>
Event instance ID ( <i>eiid</i> )	ei<integer>
Class ( <i>class</i> )	REPORTING, PERCEPTION, ASPECTUAL, I_ACTION, I_STATE, OCCURRENCE, STATE
Tense ( <i>tense</i> )	PAST, PRESENT, HYPOTHETICAL, NONE
Aspect ( <i>aspect</i> )	PERFECT, -PERFECT, FUTURE, NONE
Part of speech ( <i>pos</i> )	ADJECTIVE, NOUN, VERB, ADVERB, OTHER
Polarity ( <i>polarity</i> )	NEG, POS
Modality ( <i>modality</i> )	AHAL, BEHAR, NAHI, NONE
Certainty ( <i>certainty</i> )	CERTAIN, UNCERTAIN, UNDERSPECIFIED
Factuality ( <i>factuality</i> )	FACTUAL, COUNTERFACTUAL, NON_FACTUAL, NO_FACTUALITY_VALUE, UNDERSPECIFIED
Special cases ( <i>specialCases</i> )	CONDITIONAL_CONDITION, CONDITIONAL_MAIN CLAUSE, GENERIC, NONE

Table 6: event attributes and values in EusTimeML

The mark-up language described in the preceding sections has been used for the annotation of *EusTimeBank*, the gold standard corpus for temporal information in Basque. *EusTimeBank* is a 92-document corpus (23,000 tokens) made up of 86 news documents and 6 historical narratives. The corpus has been used for the training and

evaluation of bTime<sup>3</sup> (Salaberri 2017) and EusHeidelTime<sup>4</sup> (Altuna *et al.* 2017). Additionally, the annotated documents obtained by those tools have been used as input for KroniXa (Altuna *et al.* under revision b), a tool to build timelines from Basque texts.

News and history texts are especially rich in temporal information, as they commonly narrate past events and offer the necessary information to arrange the events along the temporal axis. Hence, their narrative nature makes these texts an interesting basis for timeline generation. For this reason, a timeline dataset for the evaluation of KroniXa has been created from *EusTimeBank* (Altuna *et al.* 2019).

## 5. DISCUSSION

The creation of EusTimeML has been the first step towards automatic temporal information extraction from Basque texts. In order to be able to compare the Basque annotated corpora and the results obtained by NLP tools for Basque with the NLP resources for other languages, comparable annotation schemes and evaluation measures should be adopted. Hence, as TimeML schemes are widely used in English, Spanish and French, building the TimeML-compliant EusTimeML has been a convenient option.

The decisions on EusTimeML have been validated by means of a set of manual annotation efforts (Altuna *et al.* 2014, 2018a, 2018b; Altuna 2018), in which inter-annotator agreement has been measured. Manual annotation analysis has shown that EusTimeML annotation guidelines are unambiguous for most of the elements, but we must note that event classification has been a major source of disagreement as annotators have considered some event classes to be virtually indiscernible in some contexts. The discussions after the agreement assessment have led to a wide consensus on EusTimeML and a consistent set of annotation guidelines has been produced (Altuna *et al.* 2016).

As our final goal is generating timelines based on the temporal information contained in texts, we have paid special attention to similar work based on TimeML annotations. In fact, the suitability of TimeML to encode temporal information for timeline building has been called into question. Ning *et al.* (2018) argue that the scarcity of intrasentential temporal relations heavily affects the event-event ordering. This

---

<sup>3</sup> Event and temporal relation extraction and classification tool.

<sup>4</sup> Time expression extraction and normalisation tool.

opinion is shared by Derczynski *et al.* (2013), as they proposed TimeML-Dense, although timeline building was not their final goal. Laparra *et al.* (2017) are also aware of the data sparsity problem for timeline building provided by TimeML annotations. They thus propose assigning the same time tag to all events a certain entity takes part in if they share the same tense, as a way to increase the number of anchored events. This partially solves the lack of temporal relations between events in the text. In Altuna *et al.* (under revision b) we have also found that some time expressions can have more than one correct normalised value in TimeML, which causes unnecessary time expression ordering problems as simultaneous events can be incorrectly placed in two different time points. For example, the quarters of the year may get different normalised values depending on whether they are referred to as quarters of a natural year or of a fiscal year.

Nonetheless, we consider that EusTimeML still offers sufficient information for timeline building. It should be taken into account that, even if bTime can only deal with a restricted set of temporal relations, experiments with KroniXa have shown very promising results, as a third of the events are correctly placed in the timelines.

Other authors have also highlighted some points in which TimeML struggles to properly encode temporal information. Ehrmann and Hagège (2009) noted that TimeML neither offered precise guidelines for time expression classification nor a clear distinction between characterisation and reference calculation annotations. According to them, a time expression such as *2 days before yesterday* should be considered a date, and *2 days* should be used to calculate its reference; TimeML proposes to annotate a duration (*2 days*) and a date (*yesterday*), instead. This same concern is shared by Bethard (2013) who proposes a scheme (SLATE) that allows machine-learning calculations.

Along the same lines, Laparra *et al.* (2018) identified the incapacity of TimeML to annotate compositional time expressions such as *Saturdays since March 6*, in which a set of dates is bounded by a determined time point. Event annotation through TimeML has also been a matter of discussion among scholars. For example, as Leeuwenberg and Moens (2019) point out, event durations cannot be explicitly tagged through TimeML, as no scheme for marking the durative (or punctual) nature of the events is provided. In spite of these flaws, TimeML is still the most widely used mark-up language for temporal information annotation.

## 6. CONCLUSIONS

EusTimeML addresses the need for a temporal information mark-up language for Basque that can deal with its language-specific features. Nevertheless, even if it contains some modifications, it is largely comparable to other TimeML-styled schemes. Adding factuality information has contributed to enlarging the amount of relevant information for timeline generation, which is our final goal.

In fact, EusTimeML has been the first step towards temporal information processing in Basque as it has been the mark-up language used for the *EusTimeBank* annotation, the corpus used for the development of the EusHeidelTime and bTime tools for temporal information extraction and normalisation. Furthermore, documents annotated following EusTimeML have also been used to generate timelines for the evaluation of KroniXa.

EusTimeML is now ready to use, although its customisability still allows for improvements and expansions. Addressing duration anchoring and increasing the amount of intrasentential temporal relations should be a goal for the TimeML community.

## REFERENCES

- Alegria, Iñaki and Kepa Sarasola. 2017. Language technology for language communities: An overview based on our experience. In Nicholas Ostler ed. *FEL XXI Alcanena 2017 Communities in Control*. Hungerford, UK: Foundation for Endangered Languages, DIDLLeS, SOAS World Languages Institute and Mercator Research Centre, 91–97.
- Altuna, Begoña. 2018. *Euskarazko denbora-egituren azterketa eta corpusaren sorrera / Analysis of Basque temporal constructions and the creation of a corpus*. Donostia: University of the Basque Country dissertation.
- Altuna, Begoña, María Jesús Aranzabe and Arantza Díaz de Ilarraza. 2014. Euskarazko denbora-egiturak. Azterketa eta etiketatze-esperimentua. *Linguamática* 6/2: 13–24.
- Altuna, Begoña, María Jesús Aranzabe and Arantza Díaz de Ilarraza. 2016. *Euskarazko denbora-egiturak etiketatzeko gidalerroak v2.0* (UPV/EHU/LSI/TR;01-2016). Donostia: University of the Basque Country.
- Altuna, Begoña, María Jesús Aranzabe and Arantza Díaz de Ilarraza. 2017. EusHeidelTime: Time expression extraction and normalisation for Basque. *Procesamiento del Lenguaje Natural* 59: 15–22.
- Altuna, Begoña, María Jesús Aranzabe and Arantza Díaz de Ilarraza. 2018a. An event factuality annotation proposal for Basque. In Andrew U. Frank, Christine Ivanovic, Francesco Mambrini, Marco Passarotti and Caroline Sporleder eds.

- Proceedings of the Second Workshop on Corpus-Based Research in the Humanities (CRH-2)*, Vol. 1. Vienna: Gerastree Proceedings, 15–24.
- Altuna, Begoña, María Jesús Aranzabe and Arantza Díaz de Ilarraza. 2018b. Adapting TimeML to Basque: Event annotation. In Alexander Gelbukh ed. *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science (LNCS)* 9624. Cham: Springer International Publishing, 565–577.
- Altuna, Begoña, María Jesús Aranzabe and Arantza Díaz de Ilarraza. 2019. EusTimeBank-TL corpora: Denbora-informaziodun testuetatik denbora-lerroetara. In Olatz Arbelaitz, Urtzi Etxeberria, Ainhoa Latatu, Miren Josu Ormaetxebarria eds. *III. Ikergazte. Nazioarteko Ikerketa Euskaraz, Giza Zientziak eta Arteak*, Vol. 1. Bilbao: Udako Euskal Unibertsitatea, 83–90.
- Altuna, Begoña, María Jesús Aranzabe and Arantza Díaz de Ilarraza. Under revision a. EusTimeBank: A corpus for temporal information processing in Basque. *Language Resources and Evaluation*. Cham: Springer International Publishing.
- Altuna, Begoña, Ander Soraluze, María Jesús Aranzabe, Olatz Arregi and Arantza Díaz de Ilarraza. Under revision b. KroniXa: Timeline creation from Basque texts. *Digital Scholarship in the Humanities*. Oxford: Oxford University Press.
- Bauer, Sandro, Stephen Clark and Thore Graepel. 2015. Learning to identify historical figures for timeline creation from Wikipedia articles. In Lucia Aiello and Daniel E. McFarland eds. *SocInfo 2014 International Workshops, Revised Selected Papers*. Barcelona, Spain: Springer, 234–243.
- Bethard, Steven. 2013. A synchronous context free grammar for time normalization. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu and Steven Bethard eds. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, USA: Association for Computational Linguistics, 821–826.
- Bittar, André. 2010. *Building a TimeBank for French: A Reference Corpus Annotated According to the ISO-TimeML Standard*. Paris: Université Paris Diderot dissertation.
- Caselli, Tommaso, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta and Irina Prodanof. 2011. Annotating events, temporal expressions and relations in Italian: The It-TimeML experience for the Ita-TimeBank. In Nancy Ide, Adam Meyers, Sameer Pradhan and Katrin Tomanek eds. *Proceedings of the 5th Linguistic Annotation Workshop*. Portland, Oregon: Association for Computational Linguistics, 143–151.
- Caselli, Tommaso and Piek Vossen. 2017. The Event StoryLine Corpus: A new benchmark for causal and temporal relation extraction. In Tommaso Caselli, Ben Miller, Marieke van Erp, Piek Vossen, Martha Palmer, Eduard Hovy, Teruko Mitamura and David Caswell eds. *Proceedings of the Events and Stories in the News Workshop*. Vancouver, Canada: Association for Computational Linguistics 77–86.
- Cassidy, Taylor, Bill McDowell, Nathanael Chambers and Steven Bethard. 2014. An annotation framework for dense event ordering. In Kristina Toutanova and Hua Wu eds. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland, USA: Association for Computational Linguistics, 501–506.
- Costa, Francisco and António Branco. 2012. TimeBankPT: a TimeML annotated corpus of Portuguese. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis eds. *Proceedings of the Eighth International Conference on Language Resources*

- and Evaluation (LREC-2012)*. Istanbul, Turkey: European Language Resources Association (ELRA), 3727–3734.
- Derczynski, Leon and Kalina Bontcheva. 2014. PHEME: veracity in digital social networks. In Harry Bunt ed. *Proceedings of the 10th Joint ACL – ISO Workshop on Interoperable Semantic Annotation (ISA)*. Reykiavik: Association for Computational Linguistics, 65–68.
- Derczynski Leon, Héctor Llorens, and Naushad UzZaman. 2013. TimeML-Strict: clarifying temporal annotation. Computing Research Repository (CoRR) abs/1304.7289. <http://arxiv.org/abs/1304.7289> (29 December, 2019.)
- Ehrmann, Maud and Caroline Hagège. 2009. Proposition de caractérisation et de typage des expressions temporelles en contexte. In Adeline Nazarenko and Thierry Poibeau eds. *Actes de la 16ème Conférence sur le Traitement Automatique des Langues Naturelles*. Senlis, France: Association pour le Traitement Automatique des Langues.
- Ferro, Lisa, Laurie Gerber, Inderjeet Mani, Beth Sundheim and George Wilson. 2003. *TIDES 2003 Standard for the Annotation of Temporal Expressions*. McLean, USA: The MITRE Corporation.
- Forăscu, Corina and Dan Tufiş. 2012. Romanian TimeBank: An annotated parallel corpus for temporal information. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis eds. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. Istanbul, Turkey: European Language Resources Association (ELRA), 3762–3766.
- Jeong, Young-Seob, Zae Myung Kim, Hyun-Woo Do, Chae-Gyun Lim and Ho-Jin Choi. 2015. Temporal information extraction from Korean texts. In Afra Alishahi and Alessandro Moschitti eds. *Proceedings of the 19th Conference on Computational Natural Language Learning, CoNLL 2015*. Beijing, China: Association for Computational Linguistics, 279–288.
- Kawai, Hideki, Adam Jatowt, Katsumi Tanaka, Kazuo Kunieda, and Keiji Yamada. 2010. Chronoseeker: Search engine for future and past events. In Dongsoo S. Kim, Sang-Wook Kim, Suk-Han Lee, Lajos Hanzo and Roslan Ismail eds. *Proceedings of the 4th International Conference on Ubiquitous Information Management and Communication, ICUIMC '10*. New York, USA: Association for Computing Machinery, 25:1–25:10.
- Kocoń, Jan and Michał Marcińczuk. 2015. Recognition of Polish temporal expressions. In Galia Angelova, Kalina Bontcheva and Ruslan Mitkov eds. *Proceedings of the Recent Advances in Natural Language Processing (RANLP 2015)*. Hissar, Bulgaria: RANLP, 282–290.
- Laparra, Egoitz, Rodrigo Agerri, Itziar Aldabe, German Rigau. 2017. Multi-lingual and cross-lingual timeline extraction. *Knowledge-Based Systems* 133, 77–89.
- Laparra, Egoitz, Dongfang Xu and Steven Bethard. 2018. From characters to time intervals: New paradigms for evaluation and neural parsing of time normalizations. *Transactions of the Association for Computational Linguistics* 6, 343–356.
- Leeuwenberg, Artuur and Francine Moens. 2019. A survey on temporal reasoning for temporal information extraction from text. *The Journal of Artificial Intelligence Research (JAIR)* 66: 341–380.
- Mani, Inderjeet and George Wilson. 2000. Robust temporal processing of news. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Hong Kong: Association for Computational Linguistics, 69–76.

- Minard, Anne-Lyse, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen and Chantal van Son. 2016. MEANTIME, the NewsReader multilingual event and time corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, H  l  ne Mazo, Asunci  n Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016)*. Portoro  , Slovenia: European Language Resources Association (ELRA), 4417–4422.
- Mostafazadeh, Nasrin, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli and James Allen. 2016. A corpus and cloze evaluation framework for deeper understanding of commonsense stories. In Kevin Knight, Ani Nenkova and Owen Rambow eds. *Proceedings of NAACL-HLT 2016*. San Diego, CA: Association for Computational Linguistics, 839–849.
- Ning, Qiang, Hao Wu and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 1318–1328.
- Otegi, Arantxa, Nerea Ezeiza, Iakes Goenaga and Gorka Labaka. 2016. A modular chain of NLP tools for Basque. In Petr Sojka, Ale   Hor  k, Ivan Kope  ek and Karel Pala eds. *Proceedings of the 19th International Conference on Text, Speech and Dialogue, TSD*. Cham: Springer, 93–100.
- Pustejovsky, James, Jos   Casta  o, Robert Ingria, Roser Saur  , Robert Gaizauskas, Andrea Setzer, Graham Katz and Dragomir Radev. 2003a. TimeML: Robust specification of event and temporal expressions in text. *New Directions in Question Answering* 3, 28–34.
- Pustejovsky, James, Patrick Hanks, Roser Saur  , Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro and Marcia Lazo. 2003b. The TimeBank Corpus. In Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery eds. *Proceedings of Corpus Linguistics 2003*. Lancaster, UK: UCREL, Lancaster University, 647–656.
- Pustejovsky, James, Marc Verhagen, Roser Saur  , Jessica Littman, Robert Gaizauskas, Graham Katz, Inderjeet Mani, Robert Knippen, Andrea Setzer. 2006. *TimeBank 1.2 LDC2006T08*. Web Download. Philadelphia: Linguistic Data Consortium. Retrieved from <https://catalog.ldc.upenn.edu/LDC2006T08>
- Pustejovsky, James, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An international standard for semantic annotation. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner and Daniel Tapias eds. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. La Valletta: Association for Computational Linguistics, 394–397.
- Radinsky, Kira and Eric Horvitz. 2013. Mining the web to predict future events. In Stefano Leonardi, Alessandro Panconesi, Paolo Ferragina and Aristides Gionis eds. *Proceedings of the sixth ACM international conference on Web search and data mining*. New York: Association for Computing Machinery, 255–264.
- Salaberri, Haritz. 2017. *Rol semantikoen etiketatzeak testuetako espazio-denbora informazioaren prozesamenduan daukan eraginaz*. Donostia: University of the Basque Country dissertation.
- Saur  , Roser. 2008. *A Factuality Profiler for Eventualities in Text*. Waltham, MA: Brandeis University dissertation.

- Saurí, Roser. 2010. *Annotating Temporal Relations in Catalan and Spanish TimeML Annotation Guidelines*. Barcelona: Barcelona Media.
- Saurí, Roser and James Pustejovsky. 2009. *Annotating Events in Catalan – TimeML Annotation Guidelines (Version TempEval-2010)*. Barcelona: Barcelona Media.
- Saurí, Roser and James Pustejovsky. 2010. *Annotating Time Expressions in Catalan – TimeML Annotation Guidelines (Version TempEval-2010)*. Barcelona: Barcelona Media.
- Saurí, Roser, Olga Batiukova and James Pustejovsky. 2009. *Annotating Events in Spanish. TimeML Annotation Guidelines (Version TempEval-2010)*. Barcelona Media.
- Saurí, Roser, Estela Saquete and James Pustejovsky. 2010. *Annotating Time Expressions in Spanish. TimeML Annotation Guidelines (Version TempEval-2010)*. Barcelona Media.
- Styler, William F., Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C. de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova and James Pustejovsky. 2014. Temporal annotation in the clinical domain. In Ellen Riloff ed. *Transactions of the Association for Computational Linguistics 2*: 143–154.
- TimeML Working Group. 2010. *TimeML Annotation Guidelines Version 1.3*. Technical report.
- Wonsever, Dina, Aiala Rosá, Marisa Malcuori and Matias Etcheverry. 2015. TEMANTEX: A markup language for Spanish temporal expressions and indicators. *Research in Computing Science* 97: 9–19.

*Corresponding author*

Begoña Altuna

Faculty of Informatics

University of the Basque Country

Manuel Lardizabal 1

20018 Donostia (Spain)

e-mail: begona.altuna@ehu,eus

received: January 2020

accepted: April 2020

# Corpus analysis of engagement discourse strategies in academic presentations

Carolina Viera<sup>a</sup> – Serena AP Williams<sup>b</sup>  
Boise State University<sup>a</sup> / United States  
Language and Heritage Institute<sup>b</sup> / United States

**Abstract** – Text analysis informed by Genre Theory (Hyon 1996) and methods in Corpus Linguistics provide the opportunity to describe language patterns that exist not only at the individual level but also in discourse communities. In this study, we investigate the discourse strategies used by novice and expert members of the academic United States (US) Spanish-speaking community to engage their audience, construct interpersonal meaning, and position themselves as expert speakers. We analyze two corpora: a specialized corpus of 32 conference presentations delivered by professors and doctoral students of Hispanic Studies, and a learner corpus of 24 in-class presentations to describe discourse patterning of social engagement expressed in text organization during presentation openings. Results indicate variation in engagement strategies between novice and expert presenters, with professors being the ones who make more use of interpersonal and interactive features to engage their audience. Our findings inform genre-based pedagogies by describing the language functions used to construct the different stages in which openings are organized. As oral presentations have been insufficiently studied (Robles Garrote 2016), this study contributes to the growing knowledge of academic oral Spanish in the United States.

**Keywords** – academic Spanish oral presentations; genre analysis; engagement; academic literacy; Spanish language teaching

## 1. INTRODUCTION<sup>1</sup>

Oral presentations are an important academic genre set comprising in-class student presentations, conference presentations, class discussions, lectures, and dissertation defenses, among others (Swales 2004; Biber 2006; Zareva 2012). Despite their importance, academic presentations had not been sufficiently studied until recently

---

<sup>1</sup> The authors would like to thank both the student participants and the conference presenter participants for access and permission to use their presentations for this study. We would also like to thank Dr. Cecilia Colombi from the University of California, Davis for granting access to her collected corpus of student presentations. Lastly, we would like to thank the anonymous reviewers for the review, requests for clarification, explanations and suggestions, and efforts towards improving our manuscript.



(Ventola 2002; Hood and Forey 2005; Seloni 2012; Robles Garrote 2016), partly because the number of available online corpora has increased (Morell and Pastor Cesteros 2018: 126). Corpora of oral language are difficult to construct and analyze in comparison to written language corpora, and this is especially evident when referring to a corpus in Spanish. Regarding learner corpora, Alonso-Ramos (2016: 7) affirms that “[t]here is no Spanish academic learner corpus such as *CALE*”, The *Corpus of Academic Learner English* for written texts.

Existing research on academic oral presentations suggests that while academic oral texts overlap in some ways with their written counterparts, distinctive features of these text types are that they showcase “research at various levels of completion, from work in progress to post-publication dissemination” (Hood and Forey 2005: 291–292), and possess a greater spontaneity than academic written texts, especially research articles or essays. Hood and Forey (2005: 292) emphasize that while “the oral performance is strongly associated with the development of a parallel written text,” the presenters must interact with an audience in the present time and place, resulting in a more interactive text (Wulff *et al.* 2009; Hyland and Jiang 2017). This highlights “the importance of interpersonal management and politeness features” (Ventola 2002: 10) in oral academic texts.

While the interactive and interpersonal character of written texts has also been studied (Hyland 2005, 2009), oral presentations require a distinct way of establishing rapport with the audience. Perhaps one of the most salient examples of this establishment of rapport is the inclusion of an interpersonal stage known as the ‘opening’ (Thompson 1994; Rowley-Jolivet and Carter-Thomas 2005; Villar 2011), a kind of preamble to the presentation content which has as its function to establish initial contact, stimulate interest, and create a dialogical setting of solidarity (Hood and Forey 2005: 292). During this stage, presenters introduce themselves, greet and acknowledge the audience, and sometimes make known the limitations of their study. In so doing, presenters utilize different discourse strategies to pique listeners’ engagement with the presentation. Openings are vital to facilitating initial understanding, which is crucial when processing information presented in real time. As this opening is not present in written texts, it constitutes a singular distinguishing element of the oral text. Presenters who include openings in their presentations show understanding of the social complexity of academic oral presentations in addition to an understanding of the

differences between oral and written texts. However, oral introductions can pose a problem, especially to novice presenters, because they are “the locus of complex pragmatic choices” (Rowley-Jolivet and Carter-Thomas 2005: 42).

In this study, we analyze the ‘opening’ in two corpora: 32 conference presentations (CPs) given by professors and doctoral students of Hispanic Studies and a learner corpus of 24 in-class presentations of learners of Spanish in the USA. We describe the language strategies used by both novices and experts to construct interaction or engagement with the audience in two different academic presentation modes. The following research questions guide our study: 1) What interactive and interpersonal discourse features are expressed in the text organization (stages) of the opening, and 2) What discourse elements are associated with expertise in academic public speaking in this context? The study is informed by Genre Theory (Flowerdew 2005; Martin and Rose 2008; Biber and Conrad 2009) and uses Corpus Linguistic methods for the data collection and analysis (Parodi 2008; Gries 2009; McEnery and Hardie 2011; Casas-Pedrosa *et al.* 2013). The study contributes to the growing field of academic oral corpus research through reporting the methodological decisions regarding annotation and tagset creation at the discourse level. The prevailing annotation of corpora is that of parts of speech while discourse-pragmatic annotation is rarer (Alonso-Ramos 2016: 14–15; Gries and Berez 2017). Consequently, the methodological decisions described in this study will be of interest to those pursuing analysis of oral language in academic settings.

Lastly, this study discusses how corpus analysis can contribute to our understanding of the Spanish academic discourse produced in academic presentations in the United States. The context of Spanish in the United States presents additional challenges to speakers in the academic community due to its multilingualism and multidialectalism. Even though research of oral academic Spanish exists in other contexts, it would be erroneous to assume that this discourse community follows the same conventions as other academic discourse communities that use Spanish. Academic oral texts in Spanish in the USA have been rarely studied. Though researchers have begun to address this sociolinguistic context (Achugar 2003, 2009; Viera 2017, 2019), there still remains a gap in knowledge with respect to the conventions of this academic community. The field of Contrastive Rhetoric has made clear that descriptions of texts within one cultural context do not always apply to those of another (Soler-Monreal *et al.*

2011). The creation of specialized corpora such as the ones discussed here allows the identification of distinctive features and discursive strategies of interaction that can later be compared with their use in other academic contexts in which Spanish is used.

## 2. GENRE-BASED AND CORPUS APPROACHES TO THE STUDY OF ACADEMIC LANGUAGE

Text analysis informed by a corpus approach provides what Flowerdew (2017) considers as an opportunity to describe language patterns that exist not only at the individual level but also in discourse communities: groups of individuals who share common goals, use and generate a set of distinctive text types (genres), develop some specific lexis and have participatory communication methods.<sup>2</sup> In our study, the academic ‘Sociorhetorical Discourse Community’ (Swales 1990) of focus consists of members who use Spanish in public places with an academic goal. Corpus analysis offers the possibility to analyze their use of the language at a larger scale than individual analyses. As Dressen-Hammouda (2012: 194) points out, these approaches at both the individual and discourse levels seek the “analysis of data toward a ‘snapshot’ view of language use, by providing a measurement of either the most frequent use or of its average use.” Studies analyzing such linguistic snapshots within a framework of Genre Theory have shown that academic discourse communities develop linguistic and discourse conventions that characterize each discipline (Burns 2001; Ciapuscio 2005; Biber and Conrad 2009). Discourse communities favor a set of textual genres or “exemplars that share similarities in structure, style, content and intended audience” (Swales 1990: 58), and that are “staged, goal-oriented social process[es]” (Martin and Rose 2007: 8). That is, production of academic texts occurs in specific contexts that determine linguistic options. With this bottom-up perspective, the analysis of texts created in these specific contexts precedes more general description of academic language patterns, thus calling for studies that add such an analysis to the more general body of knowledge.

Knowledge of genre conventions is vital to becoming an expert member in a discourse community (Swales 2004; Biber and Conrad 2009; Dressen-Hammouda 2012), and represents a challenge to the novice member of the community who has not

---

<sup>2</sup> See an extended definition, critical review and update of the concept in Swales (2016).

yet fully experienced the process of language socialization. This process implies acquiring a different style or type of discourse through participation in a new social context (Bolívar 2005; Moyano 2009; Seloni 2012). Tailoring a text for a specific, live, academic community, making necessary adjustments to the text while presenting, and interacting with a present audience are important aspects of presentations to be learned by novice members. In-class student oral presentations, which we will call ‘academic oral presentations’ (AOPs), have a pedagogical objective of adding the skill of public speaking to the student’s oral repertoire. In the academic world, conference presentations (CPs) are generally the venue in which public speaking also occurs.

Following the situational framework proposed by Biber (1994) to compare different registers, CPs and AOPs have in common the public place of communication, the planned text, and the common purpose to transfer academic knowledge. In each of these public speaking genres, presenters make their expository texts accessible to their immediate audience with whom they engage and interact. Additionally, presenters are being evaluated by their audiences, which can create language anxiety and interfere in speech production, especially at the initial part of the presentation. In a similar study, Csomay (2015: 4) compares teacher lectures and student in-class presentations and concludes that they differ in: “a) participant characteristics; b) relations among participants and c) production circumstances.” More precisely, she points out expertise and communicative purposes as the main differences between these two genres.

As part of the addressor’s epistemological stance towards the text, Parodi (2010) indicates that academic discourse should be marked by credibility and prestige. Achugar (2003) states that academic presenters should position themselves in the role of an expert. This positioning is expected in the case of the CPs where the addressor is engaged with the topic of the presentation and usually has the goal of argument in favor of an original idea. In contrast, in the AOPs, the addressor is presenting a topic that has been selected by the instructor and might not be engaging or familiar to the presenter. Expertise is achieved through participation and practice. As such, oral texts produced in AOPs and CPs represent two instances at a continuum of expertise in academic public speaking. An exploratory corpus study permits an initial approach to describing variation in expertise instantiated in the text within this particular discourse community.

### 3. METHODOLOGY

The data presented in this exploratory study come from a learner corpus and a specialized corpus. We follow a corpus-driven, “inductive approach, which progressively generalizes from the observation of data to build up the theory or rule” (Granger 2011: 13), in this case, focusing on a necessary stage of corpus research: description. The analysis of this study focuses on the opening sub-stages of the presentation (henceforth simply ‘stages’) in both AOPs and CPs to determine interpersonal and interactive (engagement) discourse features.

#### 3.1. Participants and data

We analyzed two corpora: a corpus of 32 CPs of professors and doctoral students of Spanish language and literature and a corpus of 24 groups of Spanish learners giving academic oral presentations in class. We describe the generic structure of the presentation openings of each. Table 1 describes the general characteristics of each corpus.

		AOP Corpus	CP Corpus
<b>General features</b>	Number of texts	24 groups: 91 students	32 (28 tagged)
	Number of words	43,729	74,571
	Total recorded hours	7h39	9h33
	Stage analyzed	Opening	Opening
<b>Participant features</b>	L1	English; English/Spanish bilingual	Spanish (28) English (4)
	Language level	Advanced; superior	Advanced; superior
	Education level	Upper division university	Graduate students (15) Professors (17)
<b>Linguistic features</b>		Spoken Academic Planned Monologic	Spoken Academic Planned Monologic
	Genre	In-class student presentations	Conference Presentations
<b>Textual features</b>	Topic	Sociolinguistics	Literature (18) Linguistics (14)

Table 1: Description of AOP and CP corpora

Twenty-eight CP presenters were native speakers of Spanish, and four were near-native bilingual English and Spanish speakers. Native Spanish-speaking participants originate from various Spanish-speaking countries but completed undergraduate studies at US institutions (18). Three completed bachelor’s degrees in Mexico and four in Spain while

the remaining seven participants completed their degrees in other Latin American countries. In the US, university professors are understood to have experience and membership in the academic discourse community, as they are expected to disseminate the findings of their research in public venues; therefore, we assume that professors have gained exposure to the presentational genre and are active participants of this discourse community.

The learner corpus corresponds to what Granger (2011: 11) classifies as a ‘local learner’ corpus: a smaller corpus “collected by teachers as part of their normal teaching activities and directly used as a basis for classroom materials.” Data for AOPs were collected in an upper-division Spanish class in a large Hispanic Serving Institution (HSI) in the West of the United States. The Hispanic-Serving designation is obtained when 25% of degree-seeking domestic students are classified as Hispanic. Spanish classes in the US are characterized by a mixed student population of learners of Spanish as a second language and students who learned Spanish by interaction with their family (Burgo 2017). As a result, learners of Spanish in upper division classes have varying degrees of proficiency in Spanish. In this sense, our AOP corpus is representative of the sociolinguistic learning context in the US. We acknowledge this fact proves problematic for its replicability in other Spanish-speaking communities; however, homogeneity is not a feature of naturally-occurring speech samples, especially in territories or contexts where languages are in contact. Nevertheless, we consider that the methods of this study can be replicated in other contexts in which engagement function is the focus.

### *3.2. Corpus design and task description*

Both corpora were collected between 2011–2012. The CP corpus was collected by one of the authors following all research-with-humans protocol for the protection of rights. The CP corpus data were collected in 8 different professional academic venues in different US regions.<sup>3</sup> Each presentation was part of a panel presentation of between 15 and 20 minutes in duration. Four literature CPs, which correspond to graduate students, did not include an opening stage; therefore, the total number tagged was 28.

The learner corpus was created by Cecilia Colombi, (University of California, Davis). No sociocultural or proficiency-level data accompany this AOP corpus. Since

---

<sup>3</sup> This CP is also described and used with different research purposes in Viera (2017, 2019).

the proficiency level of the presenters was not determined at the time that both the corpora had been created, the researchers assessed proficiency by listening to the presentation video recordings, applying oral proficiency interview assessment standards of the American Council of Teachers of Foreign Language (2012).<sup>4</sup>

Presentations were video-recorded, then transcribed orthographically, manually tagged, and analyzed using a corpus-based approach (McEnery and Hardy 2011). An analysis of the data followed, informed by Genre Theory (Martin and Rose 2007, 2008; Biber and Conrad 2009). Paralinguistic and visual elements were not included in the analysis because we were interested in the textual mode of communication. Both researchers checked the accuracy of all the transcriptions.

The course from which the AOP corpus was created dealt with topics related to Spanish in the United States. Students attended conference presentations, participated in pair and group discussion on each topic, and completed written exams. In addition to serving as a model, the conference sessions offered students the opportunity to learn theoretical concepts. Finally, students produced an oral presentation on one of a selection of linguistic articles related to course content. Although AOPs were group presentations, the opening is mostly delivered by one student in the group. We consider the final text a product of negotiation that reflects the linguistic options of the group. While half the groups (12) were face-to-face presentations, the other half (12) were completed by voice recording on a PowerPoint slide deck. Recorded presentations were listened to and evaluated by the members of the class. Table 2 displays the instructions excerpted from the handout provided to students for this summary task.

- 
1. Read the assigned article.
  2. Create a summary of the article.
  3. Explain the most important ideas.
  4. Use a formal register and academic vocabulary.
  5. Use a PowerPoint or other visual materials.
  6. Follow this structure:
    - a. **Introduction:** Introduce yourself. Specify the topic and objective and greet the audience. Announce your topic and goal. Make connections with the class topics.
    - b. **Development:** Cover the most important points.
    - c. **Closing:** the closing is as important as the introduction. The function is to remind the audience of the main concepts so they remain in the minds of the listeners.
- 

Table 2: AOP task instructions

---

<sup>4</sup> Please note, however, that these standards were created to rate a conversational mode of communication.

### 3.3. Corpus annotation

We created a taxonomy and tagset to identify engagement function stages in the opening. The following taxonomies served as a basis for developing a coding scheme: Rowley-Jolivet and Carter-Thomas (2005); the MICASE tagset described in Maynard and Leicher (2007) and Alsop and Nesi (2014). The corpus was manually annotated by both researchers who were familiar with both modes of presentation. Previous studies have noted the difficulty in deciding the boundaries of the tag units in the process of the corpus annotation (Alharbi and Hain 2016; Navarro and Simões 2019). To establish the cut-off points between stages, we followed Rowley-Jolivet and Carter-Thomas (2005) in using textual clues in the transcript and visual clues in the PowerPoints. Textual clues included discourse markers (i.e., *well*, *so*), and visual clues were given by the different slides of the PowerPoint and their corresponding title that acted as text organizers. We also considered pauses and gestures. Table 3 presents an example of establishing these cut-off points in a single AOP opening, along with the codes used for its annotation.

Tag	Transcript
<b>GR</b> (greeting)	<i>Buenas tardes,</i> 'Good afternoon'
[pause]	
<b>SI</b> (speaker presentation)	<i>mi nombre es X</i> 'my name is X'
[pause]	
<b>TA</b> (topic announcement)	<i>y a continuación, mis compañeras y yo tendremos a cargo el siguiente capítulo número cuatro titulado: [título del capítulo].</i> <i>Por lo cual, pido de su amable atención.</i> 'following, my classmates and I have been tasked with the following chapter number four titled: [title of chapter]. For this reason, I ask you for your kind attention'
[change of slide]	
<b>PL</b> (plan)	<i>Para iniciar con el primer tema de esta presentación, [nombre] nos expondrá</i> 'To start with the first topic of this presentation, [name] will present'
<b>DEF</b> (definition)	<i>sobre la lingüística sistémica funcional y género.</i> 'about Systemic Functional Linguistics and genre.'

Table 3. Example of discourse cues used to establish cut-off points between generic stages in one AOP opening (recordedcap4a)

Manual validation was carried out in the totality of the analyzed openings. We used a one-pass re-annotation; that is, the tagging of the corpus was repeated independently by the two different researchers. As different factors might affect intra-coder reliability (Révész 2011: 217), especially for holistic data, the researchers re-coded the data three times. We calculated inter-coder reliability following Miles and Huberman (1984) by dividing the number of agreements by the total number of decisions made. Inter-coder reliability was high (0.95), likely due to the fact that our categories are low-inference categories that “require little judgment” (Révész 2011: 212). Disagreements were resolved through discussion and where disagreement continued, elimination of the annotation from the corpus. Table 4 shows the taxonomy created for the annotation of the opening.

Functional Stage	Description and Function
<b>Greeting</b>	Speakers greet the audience
<b>Contextualizing the topic</b>	Speakers provide background information for the presentation itself or connect the presentation to a major context
<b>Topic announcement</b>	The speaker announces the topic; text functions as a text organizer (like a written title)
<b>Relevance</b>	Speakers claim the importance of the topic (centrality or need)
<b>Personal narratives</b>	Speakers present from the 1st person perspective, usually in the form of an anecdote that explains their interest in the topic of the presentation.
<b>Speaker introduction</b>	Speakers introduce themselves
<b>House-keeping</b>	Speakers pay attention to technical or organizational issues
<b>Defining the topic</b>	Speakers provide a brief explanation of the topic such as explanation, elaboration, clarification, delimiting the scope, exemplifying, reviewing, or stating the focus
<b>Thanks</b>	Speakers thank the audience or moderator
<b>Goodwill</b>	Speakers use any rhetorical strategies to achieve audience solidarity or benevolence such as self-deprecation or asking for forgiveness
<b>Humor</b>	Speakers make use of humor
<b>Presentation plan</b>	Speakers provide an outline of the organization of the presentation

Table 4: Opening structure tagset

### 3.4. Corpus analysis

A genre perspective usually entails both sequential and distributional analysis. Determining the sequential formula of the different stages is out of the scope of this study which focuses on the distribution of functional stages. Researchers working with genre analysis have proposed that a percentage of occurrence lower than 25% be considered an unstable stage of the generic structure, and values above 75% be considered prototypical, or obligatory, stages of the genre (see Navarro and Simões

2019) for a review. We classified the frequency of sub-stages as a) 25%–45%, occasional; b) 46%–74%, frequent; and c) 75%–100%, prototypical stages.

After the identification and tagging of the stages, we used the concordancer software *AntConc* 3.2.4 (Anthony 2013) to identify and quantify frequent stages. The *AntConc* Concordance Tool and Concordance Plot Tool were used to find the examples of the tags in context and the number of occurrences in the corpora. Absolute frequencies were normalized per 1,000 words. The four sub-corpora in our study were compared to determine differences in engagement discourse features: professors, graduate students, face-to-face, and recorded presentations.

### 3.5. Corpus size and representativeness

The size of the analyzed corpora is similar or larger than those discussed in the existing literature for academic oral language (see Wulff *et al.* 2009 or Robles Garrote 2016), a size that is smaller than typical written corpora because spoken data are more difficult to collect than written corpora and entail a time-consuming transcription stage. Because of the size and representativity of our corpus, our analysis applies only to our corpus: a pilot corpus that can inform a future larger corpus study. Despite its limitations, to our knowledge, no other similar corpus has been compiled in regard to spoken academic US Spanish. Therefore, the description and analysis hereby presented constitute a contribution to the field of Language for Specific Purposes as well as to genre-based approaches to teaching and learning Spanish.

## 4. RESULTS

Below, we discuss the engagement discourse strategies instantiated in the text structure of the student presentation openings and expert conference presentation openings, including the frequency of such strategies in both corpora in order to provide a description of the sub-stages used in each genre. We also compare the engagement discourse strategies of professors and graduate students within the conference presentation openings.

Table 5 reflects the number of participants that incorporated each generic stage into their presentation as well as the percentage of total participants using that stage.

The table also shows the number of individual occurrences of the feature in the corpus, indicating the frequency in each stage. The normalized frequency is indicated per 1,000 words (N=2,492).

<b>Stage Used in Opening (Word count = 2,492)</b>	<b>Participants N=28</b>	<b>Participant use (%)</b>	<b>Raw frequency</b>	<b>Normalized frequency</b>
<b>Announcing the topic</b>	19	67.86	22	8.83
<b>Contextualizing the topic</b>	18	64.29	28	11.24
<b>Defining the topic</b>	14	50.00	28	11.24
<b>Giving thanks</b>	14	50.00	20	8.03
<b>Personal narratives or personal asides</b>	11	39.29	25	10.03
<b>Explaining relevance of topic</b>	10	35.71	4	1.61
<b>Greeting audience</b>	9	32.14	9	3.61
<b>Goodwill</b>	12	42.86	17	6.82
<b>Housekeeping</b>	9	32.14	10	4.01
<b>Humor</b>	6	21.43	12	4.82
<b>Presentation plan</b>	3	10.71	2	0.80

Table 5: CPs opening structure (frequencies per 1,000 words)

Text structure analysis shows that the most frequent CP opening stages are “content-oriented and listener-oriented” in terms of Rowley-Jolivet and Carter-Thomas (2005), while the frequency of other interpersonal strategies in the CPs is occasional (less than 36%). The high-frequency stages that orient toward content include ‘topic announcement’, ‘contextualization’, and ‘defining the topic’. After announcing the topic, which is the equivalent of a title in the written mode, speakers provide background information for the presentation itself or connect the presentation with other related topics that construct shared knowledge, and succinctly define the topic by elaborating, clarifying, delimiting the scope, exemplifying, reviewing, or stating the focus of the presentation. The following examples in Table 6 illustrate the functions of the high-frequency content-oriented stages.

<b>Topic announcement</b>	The speaker announces the topic; text functions as a text organizer.	<p>1. <i>OK el título es (3LitP)</i> ‘OK the title is’</p> <p>2. <i>ah mi presentación ah tiene que ver con lo que es... (8LitE)</i> ‘uh my presentation ah has to do with what is...’</p> <p>3. <i>Bueno, yo titulé mi presentación ah (10LinP)</i> ‘Well, I titled my presentation uh’</p>
<b>Contextualization</b>	Speakers provide background information for the presentation itself or connect the presentation to a major context	<p>4. <i>trabajé con estudiantes en México así que lo que voy a presentar (10LinP)</i> ‘I worked with students in Mexico so what I’m going to present’</p> <p>5. <i>eeh un trabajo que consta de tres partes (19LitP)</i> ‘eeh a study that consists of three parts’</p>
<b>Defining the topic</b>	Speakers provide a brief explanation of the topic such as explanation, elaboration, clarification, delimiting the scope, exemplifying, reviewing, or stating the focus	<p>6. <i>en otras palabras, lo que se conoce como... (3LitP)</i> ‘in other words, what is known as...’</p> <p>7. <i>Entonces, un poco, este es justamente el entrecruce de esos dos capítulos. (10LitEH)</i> ‘So, in a way, this is the point at which these two chapters intertwine’</p> <p>8. <i>más concretamente, es una puesta en común de... (10LitE)’</i></p>

Table 6: High-frequency content-oriented opening stages

The high-frequency stages that orient toward listeners include giving thanks and personal narratives or personal asides. Table 7 displays functions of the high-frequency listener-oriented stages among the CPs.

<b>Personal narratives</b>	Speakers present from the 1st person perspective, usually in the form of an anecdote that explains their interest in the topic of the presentation.	<p>9. <i>Cuando empecé a hacer esta investigación mi mii idea era encontrarme con estudiantes recién llegados, ¿verdad? (11LinP)</i> ‘When I started this study, my, my idea was to meet with recently-arrived students, right?’</p>
<b>Giving thanks</b>	Expressing appreciation to organizers, audience-members, or other relevant individuals	<p>10. <i>Gracias por venir (1LitEM)</i> ‘Thank you for coming’</p> <p>11. <i>Gracias a los organizadores (1Lit EM)</i> ‘Thank you to the organizers’</p> <p>12. <i>Gracias Fernando (6LitPH)</i> ‘Thanks, Fernando’</p>

Table 7: High-frequency listener-oriented opening stages

Table 8 displays examples of the less frequent engagement opening stages.

<b>Relevance of the topic</b>	Speakers claim the importance of the topic (centrality or need)	13. <i>ehhh sobre todo lo que quiero llamar la atención de ustedes que trabajan con el.. la alguno...el grupo latino, que en muchos de ellos pueden llegar a ser indígenas, ¿verdad? (4LinP)</i> ‘ehhh above all what I want to call to your attention is’
<b>Goodwill</b>	Speakers use any rhetorical strategies to achieve audience solidarity or benevolence such as self-deprecation or asking for forgiveness	14. <i>...es un poquito más complicada cuando se trata de aplicar al Caribe, ¿no?; y no es que sea imposible, ¿no? pero para mí, en este momento ha sido un poquito difícil, ¿no? Entonces, este trabajo muestra esa dificultad (3LitP)</i> ‘it’s a bit more complicated when applied to the Caribbean, right?; and it’s not that it’s impossible, right? but for me, at this time it’s been a bit difficult, right? Thus, this study demonstrates that difficulty.’
<b>Humor</b>	Speakers make use of humor	15. <i>Entonces&lt;FM&gt; ...no puedo leer con las gafas (risas) (6LinE)</i> ‘So... I can’t read with my glasses (laughter)’
<b>Presentation plan</b>	Speakers provide an outline of the organization of the presentation	16. <i>la estructura de mi presentación es esta empiezo con la pregunta central luego voy a hablar un poco brevemente&lt;PL&gt; (3APNLin)</i> ‘the structure of my presentation is this: I start with the central question and then I will speak briefly’

Table 8: Less frequent interpersonal stages in CP openings

While the presenters seem to vary in terms of the selection and frequency in types of these other less frequent stages, it is important to note that when considered together, we conclude that there is an overall consistent attempt by all speakers to include interpersonal stages in the opening; on average, presenters include 5 distinct interpersonal stages in an average opening (see Figure 1 below).

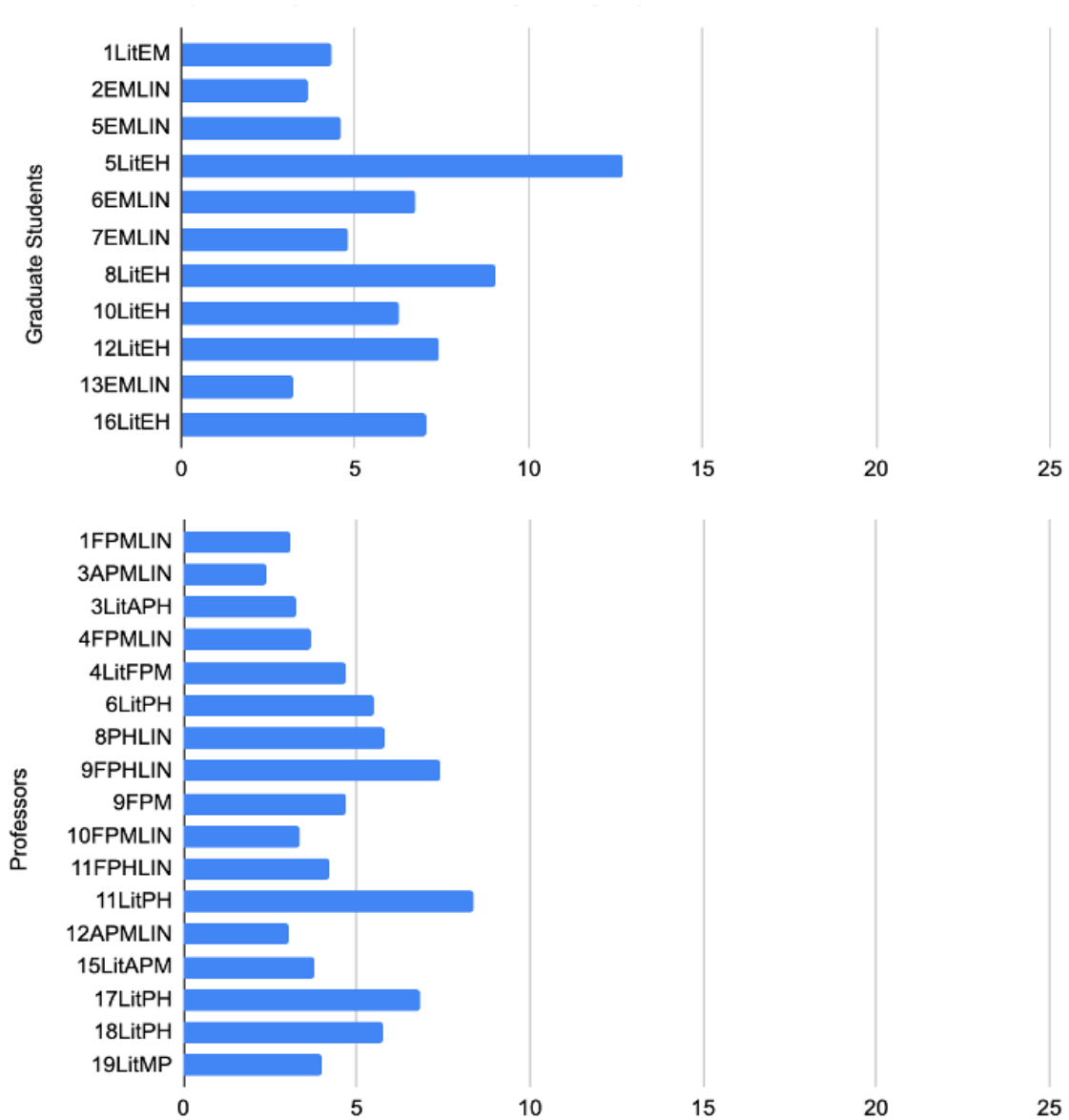


Figure 1: Number of interpersonal generic stages in CP opening (per 89 words, average CP opening length)

As graduate students may be considered peripheral members of the discourse community of academic conference presenters, we describe the frequencies of the opening stages for each group of both professors and graduate students in order to see if any distinctions in stage use exist. Table 9 below shows the stages used by the 17 professors in the CP corpus. We classified the frequency of stages as a) 25%–45%, occasional; b) 46%–70%, frequent; and c) 71%–100%, recurrent.

Stage Used	Professors N=17	Participant use %	Raw frequency	Normalized frequency
Announcing the topic	13	76.47	16	6.42
Contextualizing the topic	13	76.47	20	8.03
Personal narratives or personal asides	10	58.82	22	8.83
Defining the topic	9	52.94	12	4.82
Goodwill	9	52.94	13	5.22
Housekeeping	6	35.29	6	2.41
Giving thanks	8	47.06	13	5.22
Humor	5	29.41	9	3.61
Explaining the relevance of the topic	3	17.65	3	1.20
Greeting audience	2	11.76	2	0.80

Table 9: Professors CPs opening structure (frequencies per 1,000 words)

The most frequent content-oriented opening stages for professors include topic announcement (77%) and contextualization (77%). Graduate students also frequently utilized content-oriented opening stages (topic announcement 55%, topic definition 55%, and topic contextualization 46%); however, these frequencies do not arrive at the recurrent classification as they do for professors. For the listener-oriented opening stages, professors frequently used personal narratives or asides (59%) to engage with their audience while graduate students rarely made use of this stage (9%), preferring instead to utilize an audience greeting (64%).

In the next section, we present the findings of the text structure analysis of students' academic presentations. Table 10 reflects frequency categories for the AOP openings, with each stage within the opening showing the individual occurrences of the feature in the corpus and the number of groups that incorporated this stage into their presentation. The percentage is calculated to reflect participant usage, and the subsequent frequency categorization is indicated for each stage as well.

Stage Used	Participants (N=24)	Participant use (%)	Occurrence (hits)	Normalized Frequency
Speaker introduction	22	91.67	31	28.86
Announcing the topic	22	91.67	24	22.35
Greeting audience	20	83.33	25	23.28
Defining the topic	20	83.33	53	49.35
Contextualizing the topic	19	79.17	35	32.59
Explaining relevance of topic	4	16.67	5	4.66
Presentation plan	2	8.33	2	1.86

Table 10: Schematic structure of AOP openings (frequencies per 1,000 words)

Similar to the expert group results, students frequently include stages that are content-oriented: announcing, defining, and contextualizing the topic (see Table 11 below for examples). Indeed, when looking at the distribution of the stages, the percentage of inclusion of these stages is higher in the AOP corpus, which is an expected outcome considering that all AOPs in this corpus were collected in a similar context and students followed assignment guidelines provided by the instructor.

<b>Contextualizing the topic</b>	Speakers provide background information for the presentation itself or connect the presentation to a major context	<p>17. ...en español, en los Estados Unidos (Rec10c) [in Spanish, in the United States] la población de los hispanos en los Estados Unidos está creciendo cada día (live85) ‘the population of Hispanic in the United States is growing every day’</p> <p>18. ...una educación formal con el español puede extender el conocimiento de la lengua (Rec 9b) ‘a formal education with Spanish can extend knowledge of the language’</p>
<b>Defining the topic</b>	Speakers provide a brief explanation of the topic such as explanation, elaboration, clarification, delimiting the scope, exemplifying, reviewing, or stating the focus	<p>19. El propósito de este estudio es encontrar las cuestiones relativas a la adquisición del español (live 74) ‘The purpose of this study is to find questions relative to the acquisition of Spanish.’</p> <p>20. Y el propósito del estudio es para analizar la comparación entre el nivel um del español recibido en el aula y el porcentaje de formas, consideradas... consideradas no estándares en la producción oral de los hablantes mexicanos americanos (live80) ‘And the purpose of the study is for analyzing the comparison between a level um of Spanish received in the classroom and the percentage of forms, considered... considered nonstandard in the oral production of Mexican-American speakers.’</p>
<b>Topic announcement</b>	The speaker announces the topic; text functions as a text organizer	<p>21. vamos a hablar sobre capítulo cinco, la enseñanza del español en Nuevo México (live54) ‘we will talk about Chapter Five, the teaching of Spanish in New Mexico’</p> <p>22. y vamos a discutir el capítulo de este libro que se llama XX (Live74) ‘[and we will discuss the chapter from this book that is called XX’</p>

Table 11: Content-oriented opening stages for AOPs

Of interest for our research question on generic structure, other than greetings, students do not include interactive or interpersonal stages that were present in the expert corpus of reference, such as ‘personal narratives’, ‘humor’, ‘goodwill’ or ‘housekeeping.’ Students do include an interpersonal stage of speaker introduction which is not present in the CP corpus, and as explained in the discussion below, likely motivated by the assignment instructions in which presenters are asked explicitly to introduce themselves to their audience.

Additionally, in this student corpus, 12 of the presentations were conducted using PowerPoint narration while the other 12 were presented in a face-to-face context. Table 12 below shows that all content-oriented stages were frequent, but that there were fewer topic announcements and contextualizing stages with the face-to-face mode than with the PowerPoint narration mode.

	Topic announcement	Contextualizing	Defining
<b>Face-to-face (12)</b>	83 %	67%	83%
<b>PowerPoint with recorded narration (12)</b>	100 %	92%	83%

Table 12: Content-oriented stages of face-to-face and PowerPoint-narrated presentations

In sum, a text structure analysis of the openings of conference presentations and student presentations shows high-frequency content-oriented stages, but a difference in structural component categories and their frequency. This difference is mostly at the level of listener-oriented stages. The stages used by both experts in the reference corpus and students are summarized in Figure 2 below.

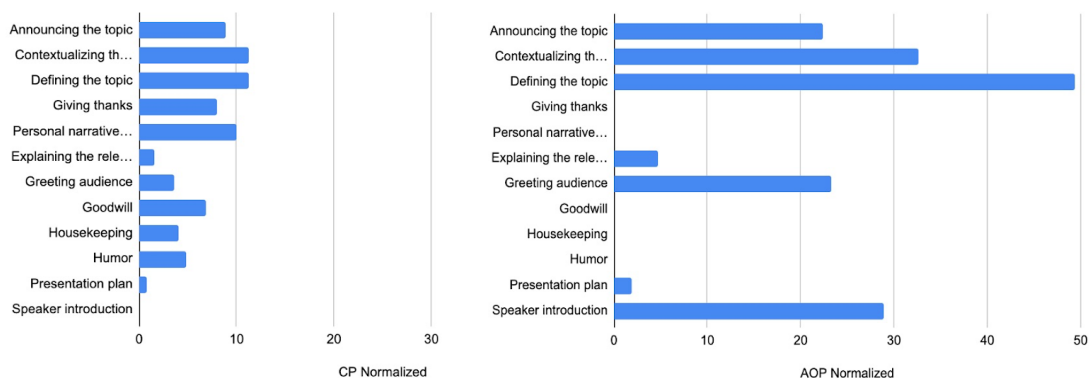


Figure 2: CP and AOP opening structure

Conference presentations use a wider variety of stages than student presentations, though both announce, contextualize, and define the topic, and to some extent, explain the role of the presentation, orienting to content. The listener-oriented stages used by students are fewer, mainly greetings and speaker introductions, while conference presenters give thanks, tell narratives, use goodwill and humor, and manage housekeeping issues.

## 5. DISCUSSION

As discussed in previous sections, the differences that we have found in the structure of CP and AOP openings reflect the variation in situational contexts of both speech events. For instance, AOP openings include a stage which is not present in the CP corpus where the speakers introduce themselves. In a conference presentation, moderators introduce the speaker, which makes this stage unnecessary. However, our analysis shows that even though these two events have different functional goals, they have structural similarities that are the result of both being public, academic speech events where speakers present cognitively-demanding information and must persuade the audience of their capabilities as valid academic communicators. Openings are crucial to achieve this interpersonal communication. Consistently, most presentations (28) analyzed in this study include an opening. With the audience in mind, they acknowledge the audience through greetings and giving thanks for their presence and make an effort to facilitate the understanding of the content of the presentation. However, at a closer look, we notice differences that show how language socialization has an impact on the academic oral texts produced by the members of a discourse community.

First, we notice that openings were absent in four graduate student conference presentations, whereas all professors included this stage. In the AOPs, students were instructed to include an opening in their presentations; therefore, the presence of openings in this corpus reflects task instructions. However, students create this stage in a very basic way, usually keeping language at the sentence level (see Table 11). For instance, in most AOPs, the topic announcement, an obligatory stage in the professor sub-corpus, is realized by stating the number of the chapter being presented or reading aloud the title of the chapter, without further defining its scope or connecting the topic of the presentation with other topics or theories discussed in class. It is important to note that the task instructions mentioned to state the goal and make connections with class

topics as part of the AOP introduction. Thus, while students include obligatory content-oriented stages in their openings—topic announcement, contextualization and defining the topic—linguistically, they construct these stages in a simpler way than professors tend to do. In doing so, they communicate less investment in engaging or facilitating the comprehension of the information they will present. One possible explanation is that students lack the language proficiency to accomplish this function, but this upper division class consisted of advanced speakers of Spanish who were able to present a complex sociolinguistic chapter in an appropriate way. Additionally, the fact that the graduate students, who were mostly native speakers of Spanish, made use of these functions, but less frequently than professors, suggests that pragmatic awareness rather than proficiency may explain the less frequent use of engaging listener-oriented stages in the opening.

Another difference between the professor corpus and the student corpus is that expert openings are divided into more stages. These additional stages consist of personal asides, housekeeping, and humor that the speaker creates in response to a specific circumstance. In the professor sub-corpus, we notice that ‘personal narratives’ and ‘personal asides’ are included in 21% of the presentations. More experienced presenters in our corpus (professors) construct a scholar identity with an active agency in the process of investigation, one in which their motivations and personal stories related to the topic or the research are equally important to the information presented. This approach is consistent with Hyland (2005: 173), who affirms that academics position themselves not as “simply producing texts that plausibly represent an external reality, but also as using language to acknowledge, construct and negotiate social relations.” In contrast, in their introductions, both graduate and undergraduate students focus on presenting information without further interaction with the audience or attempts to make the information personal or relevant. The focus on the information expressed in content-oriented stages in AOPs and lack of interaction create a text in which speakers do not position themselves in dialogue with the audience, and establish a more formal text register (Poynton 1989). In doing so, they distance themselves from the information they are presenting. This is evident when comparing face-to-face to recorded presentations in our corpus. One would expect that the first ones evidenced the presence of the audience by the inclusion of more interpersonal discourse strategies; however, the analysis shows few differences between them. It is also important to note

that making the information presented relevant at a personal level (making connections) was part of the task assignment. To the contrary, professors favor interaction and solidarity with the audience in their openings.

The professor sub-corpus analyzed here shows a discourse patterning of social engagement expressed in text organization. Interestingly, in our corpus we see a progression with respect to the importance of such strategies in relation to expertise. Regarding the differences, we observe that professors are the ones who make more use of interpersonal and interactive features of the language. Professors' openings are the site for the inclusion of personal narratives that connect the topic to the personal interests of the presenter, humor, house-keeping, and request for the benevolence or understanding of the audience if the work presented is inconclusive or a technical problem arises at the moment of the presentation. An incipient use of these strategies is seen in the graduate student sub-corpus and absent in AOPs.

By developing the taxonomy used in this research, we found that some of the categories described in previous research on academic presentations do not apply to our contexts, which to our knowledge, is a novel contribution to the field. For instance, 'presentation plan', 'explaining the relevance of the topic', and 'greeting the audience' are not prototypical stages in our expert (professor) corpus.

## 6. CONCLUSIONS

The first step in becoming aware of the main characteristics of a genre is through description because genres vary according to contexts. This study highlights that even when looking at a discrete stretch of discourse, such as the opening of a presentation, it is evident that academic discourse is complex. Our results demonstrate the importance of language description via discourse analysis and corpus research. Considering that, to our knowledge, there are no larger corpora available of oral academic Spanish in the context of the United States, this exploratory study offers a preliminary view and tools to develop more representative, larger studies.

We identified the most frequent stages of the openings in two different speech events that are part of what are collectively considered to be 'academic presentations'. We found that the number of engagement discourse strategies, which are a distinctive feature of this genre, progress with expertise. Professors create openings that facilitate

understanding of the information. They also situate themselves as active and engaged producers of the knowledge they present. In our findings, even graduate students, who are more experienced and engaged than undergraduate students, showed incipient use of such strategies. Exploring the reasons why students focus on the informational aspect of the communication escape the scope of this study but represent an interesting avenue of investigation. Our findings assert the need to investigate whether explicit instruction of academic discourse, with a genre-based approach (Schleppegrell 2004; Martin 2009) would impact language development of these markers in advanced oral proficiency in our corpora. Genres are learned through exposure, practice, and explicit teaching (Swales 2004; Fang *et al.* 2006; Antilla-Garza and Cook-Gumperz 2015) and identifying novice and expert discourse strategies provides instructors valuable information about what might be explicitly taught.

As it is the case with exploratory corpus studies, we believe that one of our main contributions can be found in the methodological decisions taken during the research process. Since we describe different genres, the analysis yielded a tagset that can be used in different contexts in future studies to analyze engagement, a crucial discourse skill for public presenters. The findings of this study, with respect to the preferred stages of an opening can also inform teaching activities designed to promote advanced literacy. Corpus informed, educational research (even a small-scale one) may contribute to our understanding of the patterning of Spanish academic discourse in specific contexts.

#### REFERENCES

- Achugar, Mariana. 2003. Academic registers in Spanish in the U.S.: A study of oral texts produced by bilingual speakers in a university graduate program. In Ana Roca and Maria Cecilia Colombi eds. *Mi Lengua: Spanish as a Heritage Language in the United States, Research and Practice*. Washington: Georgetown University Press, 213–234.
- Achugar, Mariana. 2009. Constructing a bilingual professional identity in a graduate Classroom. *Journal of Language, Identity and Education* 8/3: 65–87.
- Alharbi, Ghada and Thomas Hain. 2016. *The OpenCourseWare Metadiscourse (OCWMD) Corpus*. LREC. [http://www.lrecconf.org/proceedings/lrec2016/pdf/1085\\_Paper.pdf](http://www.lrecconf.org/proceedings/lrec2016/pdf/1085_Paper.pdf) (April 11, 2020.)
- American Council of Teachers of Foreign Language (2012). <https://www.actfl.org/> (April 11, 2020.)
- Alsop, Siân and Hilary Nesi. 2014. The pragmatic annotation of a corpus of academic lectures. In Calzolari Nicoletta, Kalid Choukri, Thierry Declerck, Hrafn Loftsson and Bente Maegaard eds. *Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation*, 1560–1563.

- Alonso-Ramos, Margarita. 2016. *Spanish Learner Corpus Research: Current Trends and Future Perspectives*. Amsterdam: John Benjamins.
- Anthony, Lawrence. 2013. *AntConc* (Version 3.2.4). <http://www.antlab.sci.waseda.ac.jp>. (April 11, 2020.)
- Antilla-Garza, Julie and Jenny Cook-Gumperz. 2015. Debating the world – choosing the word: High school debates as academic discourse preparation for bilingual students. *Linguistics and Education* 31: 276–285.
- Biber, Douglas. 1994. An analytical framework for register studies. In Biber Douglas and Edward Finegan eds. *Sociolinguistic Perspectives on Register*. Oxford: Oxford University Press, 31–58.
- Biber, Douglas. 2006. *University Language: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.
- Biber, Douglas and Susan Conrad. 2009. *Register, Genre and Style*. Cambridge: Cambridge University Press.
- Bolívar, A. 2005. Tradiciones discursivas y construcción del conocimiento en las humanidades. *Signo y Seña* 14: 67–91.
- Burgo, Clara. 2017. Meeting student needs: Integrating Spanish heritage language learners into the second language classroom. *Hispania* 100/5: 45–50.
- Burns, Anne. 2001. Analysing spoken discourse: Implications for TESOL. In Anne Burns and Caroline Coffin eds. *Analysing English in a Global Context: A Reader*. London: Routledge, 123–148.
- Casas-Pedrosa, Antonio, Jesús Fernández-Domínguez and Alejandro Alcaraz-Sintes. 2013. Introduction: The use of corpora for language teaching and learning. *Research in Corpus Linguistics* 1: 1–5.
- Ciapuscio, Guiomar. 2005. La noción de género en la Lingüística Sistémico Funcional y en la Lingüística Textual. *Revista Signos* 38: 31–48.
- Csomay, Eniko. 2015. A corpus-based analysis of linguistic variation in teacher and student presentations in university settings. In Viviana Cortés and Eniko Csomay eds. *Corpus-based Research in Applied Linguistics. In Honor of Doug Biber*. Amsterdam: John Benjamins, 1–23.
- Dressen-Hammouda, Dacia. 2012. Measuring the construction of discursual expertise through corpus-based genre analysis. In Alex Boulton, Shirley Carter-Thomas and Elizabeth Rowley-Jolivet eds. *Corpus-informed Research and Learning in ESP: Issues and Applications*. Amsterdam: John Benjamins, 193–216.
- Fang, Zhihui, Mary J. Schleppegrell and Beverly E. Cox. 2006. Understanding the language demands of schooling: Nouns in academic registers. *Journal of Literacy Research* 38/3: 247–73.
- Flowerdew, Lynne. 2005. An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies. *English for Specific Purposes* 24: 321–332.
- Flowerdew, John. 2017. Corpus-based approaches to language description for specialized academic writing. *Teach* 50/1: 90–106.
- Granger, Sylviane. 2011. How to use foreign and second language learner corpora. In Alison Mackey and Susan M. Gass eds. *Research Methods in Second Language Acquisition: A Practical Guide*. Malden: Blackwell, 7–29.
- Gries, Stefan. Th. 2009. What is Corpus Linguistics? *Language and Linguistics Compass* 3: 1–17.
- Gries, Stefan Th. and Andrea L. Berez. 2017. Linguistic annotation in/for corpus linguistics. In Nancy Ide and James Pustejovsky eds. *Handbook of Linguistic Annotation*. Dordrecht: Springer Science, 379–408.

- Hood, Susan and Gail Forey. 2005. Introducing a conference paper: Getting interpersonal with your audience. *Journal of English for Academic Purposes* 4/4: 291–306.
- Hyland, Ken. 2005. Stance and engagement: A model of interaction in academic discourse. *Discourse Studies* 7/2: 173–192.
- Hyland, Ken. 2009. Corpus informed discourse analysis: The case of academic engagement. In Maggie Charles and Susan Hunston eds. *Academic Writing: At the Interface of Corpus and Discourse*. London: Continuum, 110–128.
- Hyland, Ken and Feng Jiang. 2017. Is academic writing becoming more informal? *English for Specific Purposes* 45: 40–51.
- Hyon, Sunny. 1996. Genre in three traditions: Implications for ESL. *TESOL Quarterly* 30/4: 693–722.
- Martin, James. 2009. Genre and language learning: A social semiotic perspective. *Linguistics and Education* 20/1: 10–21.
- Martin, James R. and David Rose. 2007. *Working with Discourse: Meaning beyond the Clause*. London: Continuum.
- Martin, James R. and David. Rose 2008. *Genre Relations: Mapping Culture*. London: Equinox Publishing.
- Maynard, Carson and Sheryl Leicher. 2007. Pragmatic annotation of an academic spoken corpus for pedagogical purposes. In Eileen Fitzpatrick ed. *Corpus Linguistics beyond the Word: Corpus Research from Phrase to Discourse*. Amsterdam: Rodopi, 107–115.
- McEnery, Tony and Andrew Hardie. 2011. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
- Miles, Matthew B. and Michael Huberman. 1984. *Qualitative Data Analysis: A Sourcebook of New Methods*. Beverly Hills: Sage.
- Morell, Teresa and Susana Pastor Cesteros. 2018. Multimodal communication in academic oral presentations by L2 Spanish students. *Journal of Spanish Language Teaching* 5/2: 125–138.
- Moyano, Estela Inés. 2009. Negotiating genre: Lecturer's awareness in genre across the curriculum Project at the university level. In Charles Bazerman, Adair Bonini and Débora Figueiredo eds. *Genre in a Changing World*. Indiana: Parlor Press and WAC Clearinghouse, 442–464.
- Navarro, Federico and Alex Simões. 2019. Potencial de estructura genérica en tesis de ingeniería eléctrica: Contrastes entre lenguas y niveles educativos. *Revista Signos* 52: 306–329.
- Parodi, Giovanni. 2008. Lingüística de corpus: Una introducción al ámbito. *Revista de Lingüística Teórica y Aplicada* 46/1: 93–119.
- Parodi, Giovanni. 2010. *Academic and Professional Discourse Genres in Spanish*. Amsterdam: John Benjamins.
- Poynton, Cate. 1989. *Language and Gender: Making the Difference*. Oxford: Oxford University Press.
- Révész, Andrea. 2011. Coding second language data validly and reliably. In Alison Mackey and Susan M. Gass eds. *Research Methods in Second Language Acquisition: A Practical Guide*. Malden: Blackwell, 201–221.
- Robles Garrote, Pilar. 2016. Aportaciones de la lingüística de corpus al estudio de la conferencia como género académico de divulgación científica. *Chimera: Romance Corpora and Linguistic Studies* 3: 1–21.

- Rowley-Jolivet, Elizabeth and Shirley Carter-Thomas. 2005. The rhetoric of conference presentation introductions: Context, argument and interaction. *International Journal of Applied Linguistics* 15/1: 45–70.
- Seloni, Lisy. 2012. Academic literacy socialization of first year doctoral students in the USA: A micro-ethnographic perspective. *English for Specific Purposes* 31/1: 47–59.
- Schleppegrell, Mary. J. 2004. *The Language of Schooling: A Functional Linguistics Perspective*. Mahwah: Erlbaum.
- Soler-Monreal, Carmen, María Carbonell-Olivares and Luz Gil-Salom. 2011. A contrastive study of the rhetorical organisation of English and Spanish PhD thesis introductions. *English for Specific Purposes* 30/1: 4–17.
- Swales, John. M. 1990. The concept of discourse communities. In John M. Swales ed. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press, 21–32.
- Swales, John. M. 2004. *Research Genres: Explorations and Applications*. Cambridge: Cambridge University Press.
- Swales, John. M. 2016. Reflections on the concept of discourse community. *ASp* 69: 7–19.
- Thompson, Susan. 1994. Frameworks and contexts: A genre-based approach to analysing lecture introductions. *English for Specific Purposes* 13/2: 171–186.
- Ventola, Eija. 2002. Why and what kind of focus on conference presentations? In Eija Ventola, Celia Shalom and Susan Thompson eds. *The Language of Conferencing*. Bern: Peter Lang, 15–50.
- Viera, Carolina. 2017. Genre and register variation: Academic conference presentations in Spanish in the United States. In Juan Colomina-Almiñana ed. *Contemporary Advances in Theoretical and Applied Spanish Linguistic Variation*. Columbus: Ohio State University Press, 148–161.
- Viera, Carolina. 2019. La informalidad como recurso en el español académico de los Estados Unidos. In Gregory L. Thompson and Scott Alvord eds. *Contact, Community, and Connections: Current Approaches to Spanish in Multilingual Populations*. Wilmington: Vernon Press, 59–82.
- Villar, Claudia. 2011. Las presentaciones académicas orales de los estudiantes alemanes de E/LE. Del discurso monológico al dialógico. *Revista Nebrija de Lingüística Aplicada* 10/5: 130–172.
- Wulff, Stephanie, John. M. Swales and Kristen Keller. 2009. ‘We have about seven minutes for questions’: The discussion sessions from a specialized conference. *English for Specific Purposes* 28/2: 79–92.
- Zareva, Alla. 2012. Self-mention and the projection of multiple identity roles in TESOL graduate student presentations: The influence of the written academic genres. *English for Specific Purposes* 32/2: 72–83.

*Corresponding author*

Carolina Viera

Boise State University

World Languages Departament

1910 University Drive

83725 Boise, Idaho

United States

e-mail: [carolinaviera@boisestate.edu](mailto:carolinaviera@boisestate.edu)

received: January 2020

accepted: April 2020

# The TAGFACT annotator and editor: A versatile tool

Ana Fernández-Montraveta<sup>a</sup> – Hortènsia Curell<sup>a</sup> – Glòria Vázquez<sup>b</sup> – Irene Castellón<sup>c</sup>  
Universitat Autònoma de Barcelona<sup>a</sup> / Spain  
Universitat de Lleida<sup>b</sup> / Spain  
Universitat de Barcelona<sup>c</sup> / Spain

**Abstract** – The multifunctional tool this paper presents has been developed within the TAGFACT project, a project that aims to automate the annotation of factuality –understood as the degree of commitment with which the writer presents situations– in Spanish journalistic texts. In what follows, the tool, which allows the compilation of the texts and the manual annotation of predicates, is described. The corpus created using it has been extracted in groups of three pieces of news covering the same event from newspapers with different ideologies (left wing, right wing and centrist). It is made up of 176 different pieces of news, containing 1,359 sentences and 46,947 words. The tool has been used so far to manually annotate a section of the ‘Gold Standard’ (approximately 10,000 words). It has proved to be versatile in that it allows for both the creation and management of corpora and corpus annotation, using any tags the user wants depending on the purpose of each corpus.

**Keywords** – annotation tool; corpus creation; corpus edition; Spanish journalistic texts

## 1. INTRODUCTION

The categorization of events with respect to their factual status is an area of growing interest in the field of Corpus Linguistics and Natural Language Processing. In recent years, several projects dealing with the annotation of corpora, either manual or automatic, with this type of information have been developed. So far, the most common approach has been the annotation of the degree of certainty with which the author of a message presents an event (Saurí 2008).

The objective of our project (TAGFACT), which is two years into its development, is to create a system for the automatic annotation of the degree of certainty implicit in the situations narrated in Spanish journalistic texts, an annotation

solely grounded on linguistic knowledge (Alonso *et al.* 2018).<sup>1</sup> In Spanish, this issue has not been dealt with in much depth, and what little has been done is based primarily on statistical processes (Wonsever *et al.* 2016).

One of the first steps in our project was the creation of a corpus of Spanish journalistic texts (the TAGFACT corpus) and then a portion of this corpus, which will constitute the ‘Gold Standard’, is being annotated manually. In order to perform these two tasks, the tool presented here was created. Before presenting the tool, it is necessary to describe briefly the main aspects of the project. Thus, Section 2 presents a brief state of the art and sets the framework for our annotation scheme –described in Vázquez and Fernández-Montraveta (in press)– required to fully understand the tool. Section 3 describes the design of the corpus and the ‘Gold Standard’ and, finally, Section 4 presents the tool and how it can be used to collect corpora and carry out the manual annotation.

## 2. THE ANNOTATION OF FACTUALITY

One of the groundbreakers in the annotation of factuality in texts is *FactBank* (Saurí and Pustejovsky 2009), which constitutes an innovative proposal for the representation of this semantic category in English. *FactBank* contains 9,488 events manually annotated with factuality information, and it also takes into account the source of information.

Various authors have drawn on Saurí and Pustejovsky (2009) for the annotation of different corpora, with the factuality values established using exclusively information from the text. In this respect, some projects worth mentioning are Diab *et al.* (2009), Soni *et al.* (2014), Tonelli *et al.* (2014), van Son *et al.* (2014) and Lee *et al.* (2015) for English; Matsuyoshi *et al.* (2010) and Narita *et al.* (2013) for Japanese; Minard *et al.* (2016) for Italian; Wonsever *et al.* (2016) for Spanish; and Velupillai (2011) for Swedish. Other authors, contrary to the framework used in *FactBank*, have considered factuality as linked to the knowledge of the world (Marneffe *et al.* 2012).

In our project, we basically follow Saurí and Pustejovsky (2009) and Diab *et al.* (2009), although we propose some innovations in the annotation scheme. The first decision is whether a predicate will be annotated or not. If it is decided not to, the

---

<sup>1</sup> The authors would like to acknowledge the support from the Ministerio de Economía, Industria y Competitividad: Research Project ‘Del texto al conocimiento. Factualidad y grados de certeza en español –TAGFACT’ (Grant number FFI2017–84008– P).

predicate is disregarded altogether. If it is annotated, four categories are used: ‘Polarity’, ‘Degree of commitment’, ‘Time’ and ‘Dynamicity.’ Following Diab *et al.* (2009), we prefer the term ‘commitment’ rather than ‘certainty’ (Saurí and Pustejovsky 2009), since it reflects better that we are describing the author’s view of the event.

Regarding ‘Time’ –following van Son *et al.* (2014), Wonsever *et al.* (2016) and Matsuyoshi *et al.* (2010)– we assign one of the following values: ‘Present’, ‘Past’ or ‘Future.’ Future situations are different from present and past ones, since they can never denote facts that have happened at the point of narration. It could be argued, hence, that certainty does not apply to them. However, we claim that the writer can present a future situation with commitment or with lack of it, and this is one of the innovations of our project. Another important novelty is the inclusion of ‘Dynamicity’, in which not only do we distinguish between states and events, but we also provide a fine-grained annotation of states. Following Tonelli *et al.* (2014) and van Son *et al.* (2014), we treat absolute truths –as in *The Earth is round*– and habits –as in *In our country we usually have lunch at 2 p.m.*– differently from other types of stative situations. In the former, commitment is always stronger since they represent knowledge commonly agreed upon by a community. The latter are of interest because they always include more than one situation and some of the events are in the past, some in the present and some in the future.

Furthermore, following Saurí (2008), van Son *et al.* (2014) and Prabhakaran *et al.* (2015), among others, it was decided to signal the authorship or source of each commitment, considering that there is not such a thing as ‘reality’ and that facts are always narrated from a given perspective. In addition, our annotation includes the predicate and all the entities involved in it, since we believe that a fact necessarily contains all the participants in the situation.

Regarding the automation of the annotation process, the systems currently available follow two different approaches: those using machine-learning techniques and those based, at least partially, on linguistic information. Among the former, Mullick *et al.* (2019) present the development of a deep neural network based on the ‘Factuality Judgment Model’, while Huang *et al.* (2019) use ‘Bi-directional Long Short-Term Memory’ (BiLSTM), that is, neural networks to learn contextual information about the event in sentences. The latter consider that annotating factuality at sentence level provides an incomplete picture and their unit of analysis is the document.

On the other hand, *De Facto* (Saurí 2008) automates part of the annotation of factuality using knowledge extracted from a corpus, that is, linguistic information. It is not fully automatic, though, since some knowledge modules were created manually. Different kinds of automatic tools have also been developed for various languages, for example Minard *et al.* (2006), Narita *et al.* (2013) and Lee *et al.* (2015). In TAGFACT, only linguistic information is used to tag factuality.

As regards edition and annotation tools, there are various tools available nowadays. Some of these are designed with a general purpose and are tools for the creation, annotation and edition of corpora at different levels –*UAM Corpus Tool* (O'Donnell 2008). Some of them include the functionality of defining the categories ('Tagset'), together with the possibility to annotate at different linguistic levels. Other tools incorporate the automatic treatment of certain aspects, such as the segmentation into sentences (sentence split) or the identification of words (tokens)– *ANNIS* (Krause and Zeldes 2016). Still other annotators aim at a specific type of corpus or level of analysis, such as *Knowtator* (Ogren 2006) and *DART* (Weisser 2006). Most annotators include the production of output in XML format, the possibility of conducting complex searches and statistical tools. Our tool is versatile since, on the one hand, it allows the user to organize and manage the texts compiled for the corpora, to interact with the database created (in MySQL), to extract the final data in XML and to create tabs, while still permitting the automatic processes of tagging and parsing.

### 3. THE TAGFACT CORPUS

The TAGFACT corpus includes news articles from several Spanish newspapers. Specifically, three pieces of news, narrating the same event, were collected from newspapers with different ideologies: right wing (*La Razón*), left wing (*El Diario*) and centrist (*El Periódico*).<sup>2</sup> This will eventually allow the analysis of the author's stance and the role of ideology in journalism. The articles were mainly chosen from two genres: politics and sports. Those two genres offer the possibility of finding news and describing facts more than opinions (as opposed to, for example, op-ed columns). As for politics, since the newspapers represent clearly different ideologies, it is to be expected that the perspective from which certain events are narrated will be distinct. In sports, the

---

<sup>2</sup> When it was not possible to find one of the pieces in these media, articles were taken from another one with similar characteristics (*ABC*, *Público*, *La Vanguardia* or *20 Minutos*, among others).

well-known rivalry between the football teams Barcelona and Madrid will guarantee varied points of view. At present, the corpus includes 176 different pieces of news, containing 1,359 sentences and 46,947 words.

For each piece of news, metadata is saved in order to facilitate the access to the information about the source: name of the newspaper, section, date, author, news URL and geographical location. The data is structured in several fields, following the structure of the newspaper article. Any extra information, which is an informative part of the piece, is also included in this structure: namely the title and subtitle, the text and images and TWITTER comments, as shown in Figure 1. At present, we are manually annotating a part of the corpus, which will constitute the ‘Gold Standard.’ The total volume of words in the ‘Gold Standard’ corpus is approximately 10,000.

<b>Group</b> 1 - Máster Cifuentes		<b>Item creator</b>	
<b>Newspaper</b> 20 Minutos		<b>Section</b> Technology	
<b>Date</b> 19/03/2019	<b>Author</b> Judith Vives		
<b>Title</b> Así alimenta Youtube las teorías que afirman que la Tierra es plana			
<b>Subtitle</b> Una investigación sugiere que la plataforma de vídeos ayuda a los teóricos de la conspiración tierraplanista			
<b>Text</b> Youtube podría estar desempeñando un papel importante para convencer a algunas personas de que la Tierra es plana. Así lo sugiere un estudio de la Texas Tech University que se ha basado en realizadas a personas que han participado en conferencias sobre el tema. Las entrevistas realizadas a estas personas demuestran, según el estudio de la Texas Tech University, que la mayoría basan sus creencias en los vídeos que han visto en Youtube. Estos videos trat demuestran que la tierra no es redonda.			

Figure 1: Partial structure of the data for each item in the corpus

#### 4. THE TAGFACT TOOL

A multi-purpose tool was created to compile and annotate corpora. This tool has two main functionalities: first, to compile and manage large collections of text and, second, to facilitate the annotation of any specific corpus. The first functionality allows corpus creation, edition and management through a highly user-friendly interface. Data is collected and saved in a MySQL Database. The interface offers the possibility of querying the database and allows the downloading of all the information in either Excel or XML formats.

#### 4.1. Corpus creation, edition and management

The corpus creation tool permits the compilation of one or several collections of texts, each of which can be saved as an independent database and can be independently named, edited and modified. In addition, once a particular corpus has been collected, the tool allows for the edition, modification and management of each item in the collection, regardless of whether it has already been annotated.

The tool includes a default administrator that has the capacity to add any number of users able to interact with the database in different ways, depending on the role assigned to them (administrator or annotator). Besides assigning roles to the collaborators in the project, the administrator is in charge of assigning the sections that each annotator has to deal with, and can view all the annotations, as shown in Figure 2.

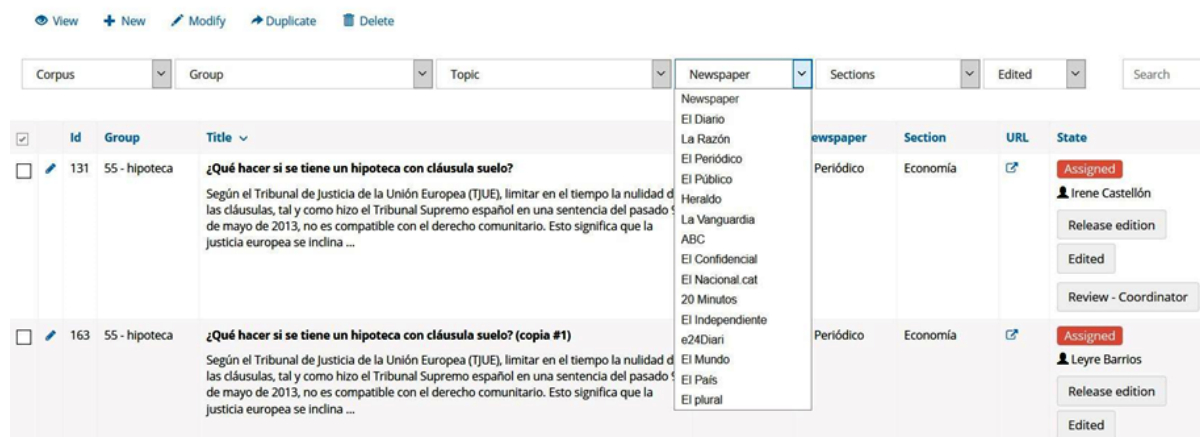


Figure 2: News items and newspapers in the corpus

The tool is connected to a parser that analyzes the texts. The administrator sends the text to the parser and has access to the pieces of news at all stages in the process. Annotators have access exclusively to items assigned to them by the administrator, as illustrated in Figure 3.

Id	Title	Text	Date	Newspaper	Section	URL	State
87	Los bomberos de Lesbos: "Estamos condenados a salvar vidas, no por salvarlas"	Manuel Blanco, Enrique Rodríguez y Julio Latorre esperan poner fin este lunes a una pesadilla "surrealista" que comenzó en enero de 2016. Estos tres bomberos sevillanos miembros de la asociación Proem-Aid, que participaban de forma voluntaria en los peores días de la crisis de refugiados en Grecia,	06/05/2018	El Periódico	Sociedad		Assigned Gloria Vázquez Release edition Edited Review - Coordinator
4	Los bomberos españoles que ayudaron a inmigrantes en Grecia llegan a juicio: "Nuestra condena sería un aviso a navegantes"	Onio Reina recuerda como si fuera ayer el momento en el que pisó por primera vez la arena de las playas de Lesbos. Era diciembre de 2015 y este bombero sevillano acababa de llegar a la isla griega junto a otros tres compañeros tras un largo trayecto en coche desde España. Uno de los vecinos les ...	06/05/2018	El Diario	Política		Processed 14/01/2019 12:24:56 Assign to...
5	Tres bomberos españoles, juzgados en Grecia por tráfico de personas	Manuel Blanco, José Enrique Rodríguez y Julio Latorre, los tres bomberos españoles detenidos en 2016 en Lesbos por la Guardia Costera Griega cuando realizaban tareas de rescate de refugiados con la ONG Proem-AID, han llegado esta mañana al tribunal que les va a juzgar por un presunto delito de ...	07/05/2018	La Razón	Política		Processed 14/01/2019 12:25:43 Assign to...

Figure 3: Corpus management

Texts undergo several stages through the process: 'Initial', 'Pending', 'Processed', 'Assigned' and 'Edited.' 'Initial' state means that an item in a collection has been introduced and documented (metadata); then, it becomes 'Pending.' It is subsequently sent to the parser, which sends it back as 'Processed', that is, segmented into sentences and with the list of predicates and their corresponding arguments, as shown in Figure 4. When the administrator assigns an item to an annotator, the stage will change to 'Assigned', and the process of annotation can start. While annotating the factual values, the syntactic structure can be corrected or 'Edited', if required.

**Predicates:**

El candidato de el PSOE a la Presidencia\_de\_el\_Gobierno , Pedro\_Sánchez , llegó el jueves a el debate sobre la moción de censura contra Mariano\_Rajoy con el acuerdo previo con el PNV de respetar los Presupuestos aprobados por el Gobierno y , con ello , garantizar se el apoyo mayoritario de la Cámara .

+ New ← Back

Trig	Categories	Arguments	Trig/Voice	Main	Probl
llegó (t1.12) *	Applies *	El candidato de el PSOE a la Presidencia_de_el_Gobierno , Pedro_Sánchez ,	A 11	✓	✗
	Past *	a el debate sobre la moción de censura contra Mariano_Rajoy	A 10		
	Commitment	el jueves	A 2		
	Positive *	con el acuerdo previo con el PNV de respetar los Presupuestos aprobados por el Gobierno y , con ello , garantizar se el apoyo mayoritario de la Cámara	A 28		
	Event *				

Figure 4: Editor interface – predicate *llegó* 'arrived'

#### 4.2. Corpus tagging

As mentioned above, the first step in the annotation process is to send the text to an external analyzer. The most complete tools for Spanish, in terms of the different levels of analysis provided, were considered in order to choose the most adequate parser for our project. According to Soroa *et al.* (2017), these are *Freeling* (Padró and Stanilovsky 2012) and *Ixa Pipe* (Agerri *et al.* 2014). Both offer a level of document representation

and resolve co-referencing, in addition to providing a morphological and syntactic analysis. *Freeling* is the parser that has been more thoroughly evaluated and has obtained an optimal index for syntactic parsing, more specifically 84% of accuracy in the analysis of dependencies (Lloberes *et al.* 2015). No evaluation of syntactic performance of the *Ixa Pipe* parser was found, even though other levels, such as co-reference, have been assessed, achieving 55% of accuracy. In addition, some tests using our corpus were performed and the final decision was to use *Freeling* as a basic working tool.

An important problem presented by the *Freeling* output is the identification of predicates, more specifically, the recognition of eventive nouns. Regretfully, the recognition of this type of element does not seem to work well, as shown in (1), and the decision to deactivate all eventive nouns has been made.

- (1) *Esto significa que la justicia europea se inclina porque la banca tenga que devolver a sus clientes lo cobrado de más.*

‘This means that the European justice favors that the banks return the money overcharged to their clients.’

The analysis of compound verbs –both complex tenses and verb periphrases– is problematic in *Freeling* as well, since it separates the verbs in a complex as two (or more) independent predicates. This problem has been easily overcome with a pre-process that rewrites them as one single predicate. In the case of complex tenses, a simple rule identifies those structures in which *haber* ‘have’ is followed by a past participle. As for verb periphrases, another rule identifies periphrastic verbs and unites them with the corresponding main verb. The only problematic issues that cannot be solved automatically are the cases in which there is one or more lexical items placed between the auxiliary and the main verb, as shown in (2). These cases are dealt with manually.

- (2) ... *ya estaba en el agua rescatando a inmigrantes que partían de la costa.*

‘... he was already in the water rescuing immigrants leaving the coast...’

#### 4.3. Corpus annotation

Once an item of the collection has been returned from the parser one can proceed to the manual annotation. The user can validate the structure sent by the parser through the

interface and categorize each predicate regarding its factual status, according to the categories proposed in the scheme (Section 2). In our project, we propose four categories, but the number can be increased or decreased by the administrator. Figure 5 shows the first layer of annotation.



Categories of predicate			
<a href="#">View</a> <a href="#">+ New</a> <a href="#">✎ Modify</a> <a href="#">🗑 Delete</a>			
<input checked="" type="checkbox"/>	Name	Default	Order ▾
<input type="checkbox"/>	<a href="#">✎</a> NA (Does not apply)	✖	↑ 10 ↓
<input type="checkbox"/>	<a href="#">✎</a> Applies	✔	↑ 20 ↓
<input type="checkbox"/>	<a href="#">✎</a> Eventive noun	✖	↑ 30 ↓
<input type="checkbox"/>	<a href="#">✎</a> Error: no predicate	✖	↑ 40 ↓

Figure 5: Tag creation and management

This layer allows the annotator to make the first decision: whether the annotation of the factual status of an event is relevant (‘Applies’ vs. ‘(NA) Does not apply’). A predicate is only labeled as ‘Does not apply’ when the clause describes a wish or a conjecture, such as *deseara* ‘wished’, as illustrated in (3).

- (3) ... *podría, si lo deseara, poner fin a la investigación o incluso ejercer su poder de perdón.*  
 ‘... he could, if he wished, end the investigation or even exercise his power of forgiveness.’

The other options in this layer of annotation are: ‘Eventive noun’ –as explained above, these nouns are not annotated at the current stage of the project– and ‘Error: no predicate’ for words identified as eventive which are, in fact, not eventive, as shown in (4).

- (4) *Hace unos meses Mongolia lanzó una campaña de apoyo para recaudar fondos.*  
 ‘A few months ago Mongolia launched a fundraising support campaign.’

Finally, if the annotators consider that the predicate has to be tagged with respect to factuality, they use ‘Applies.’ The next step is to determine the following aspects: the time referred to by the predicate (present, past or future), the degree of the writer’s commitment towards the truth or falsehood of the predicate (‘Commitment’ or ‘Non-commitment’), polarity (‘Positive’ or ‘Negative’) and, finally, dynamicity (‘Event’, ‘Mental Predicate’ or ‘Property’).

Regarding temporal information, tense and time do not always correspond, as can be seen in (5), where a present tense indicates a past time.

- (5) *Landrum alerta que el algoritmo que sugiere nuevos vídeos a las personas que buscan información sobre este tema les acaba llevando a un pozo de información incorrecta.*

‘Landrum warns that the algorithm that suggests new videos to people looking for information on this topic ends up leading them to a reservoir of incorrect information.’

Future situations are labeled differently from uncertain past and present situations because they are radically different in nature. Only in the first case is uncertainty absolute since future situations have not happened yet. The author can only express (non)-commitment towards the possibility of situations happening in the future.

Regarding polarity, one value, ‘Positive’ or ‘Negative’, is applied to the whole sentence, as in (6). Although polarity can have different scopes, at the present stage the tool only allows to assign polarity to the whole predicate. This is a limitation of the project that can be addressed in the future.

- (6) *Estos vídeos tratan de mostrar evidencias que demuestren que la tierra no es redonda.*

‘These videos try to present evidence that proves that the Earth is not round.’

The ‘Dynamicity’ tag accounts for the internal structure of predicates, differentiating between stative, dynamic situations (events) and mental processes. Stative situations express properties of individuals or events, whereas events refer to actions or processes that happen in the world and have the capacity to modify it, as illustrated in (7). Mental predicates describe cognitive processes, as shown in (8). As for stative situations, if the property refers to individuals (both people and objects), the tag used is ‘Non-eventive Property’, and when it refers to events, ‘Property Event’, as can be seen in (9) and (10) respectively. Finally, ‘Property-Absolute Truth’ is used for properties considered as such by culture or scientific proof, as shown in (11).

- (7) *El estudio se ha realizado a partir de las entrevistas con 30 asistentes a dos conferencias sobre teorías de la Tierra plana.*

‘The study was carried out based on interviews to 30 attendees at two conferences about flat Earth theories.’

(8) *En los últimos tiempos han proliferado las personas que no aceptan la idea de que el planeta Tierra es redondo.*

‘In recent times, people who do not accept the idea that planet Earth is round have proliferated.’

(9) *No es un ataque político, es un ataque personal.*

‘It is not a political attack; it is a personal attack.’

(10) *Landrum alerta que el algoritmo que sugiere nuevos vídeos a las personas...*

‘Landrum alerts that the algorithm that suggests new videos to people...’

(11) *.... el planeta Tierra es redondo*

‘... planet Earth is round.’

As pointed out above, the final goal of the project is to develop an automatic tool for the recognition and annotation of factuality. To this aim, the editor permits the annotation of linguistic cues (triggers), either morphological or lexical, that justify the choice of a tag so that they can be used in the automatic tool. For example, in (12), the clause containing the verb *tiene* ‘has’ is annotated, whereas the conditional clause is not since it is a condition. The trigger for its interpretation is the word *si* ‘if.’ Similarly, in (13), the verb *explica* ‘explains’ would be annotated as a trigger for the interpretation of the clause as a commitment for two reasons: first, the verb tense used (present indicative) and, second, the semantic class that the verb belongs to (verb of communication).

(12) *Pero al margen de ese posible acuerdo existen algunas opciones en función de si se tiene la cláusula suelo en la hipoteca...*

‘But apart from this possible agreement there exist some options depending on whether your mortgage has a base clause...’

(13) *... explica Óscar Serrano, abogado del “Col·lectiu Ronda”, exigiendo la devolución íntegra y retroactiva de los intereses pagados de más.*

‘... explains Óscar Serrano, a lawyer with the “Col·lectiu Ronda”, demanding the full and retroactive return of interest paid in excess.’

The possibility of annotating any relevant voices in the narration other than the writer’s is also considered, because they might modify the interpretation of the event. The author of the piece of news is always considered the main narrator, presenting events and situations from a particular perspective. When the main author provides the name of a different narrator, as in (14), and explicitly states the source of the information, it is understood that the author is somehow moving away from it.

(14) *Algunas de las personas consultadas aseguran que al principio solo miraban los videos para criticarlos...*

‘Some of the people consulted claim that at first they only watched the videos to criticize them...’

The tool provides a field where any problems encountered during the annotation process can be recorded and then discussed before the next stage in the project, as shown in Figure 6. Keeping a log of doubts allows the creation of lists of problematic configurations with the view of the systematization of the annotation.

<input type="checkbox"/>		3	Pero , ahora se abre un proceso que aún no está de el todo claro : ¿ devolverán de oficio los bancos el dinero cobrado de más o los afectados tendrán que acudir a los tribunales para reclamar lo ?		A 40	8
<input type="checkbox"/>		4	La sentencia europea no deja claro cómo tiene que aplicar se la resolución .	1	A 14	3
<input type="checkbox"/>		5	Puede que esta duda se resuelva políticamente si el Gobierno y la oposición se ponen de acuerdo sobre cómo hacer que los bancos devuelvan el dinero .		A 27	7
<input type="checkbox"/>		6	Pero a el margen de ese posible acuerdo existe algunas opciones en función de si se tiene la cláusula suelo en la hipoteca y no se ha reclamado ante los tribunales , si se ha planteado una demanda que ya ha sido resuelta o si se ha alcanzado un acuerdo con la entidad financiera para cambiar la hipoteca de tipo variable con suelo a tipo fijo .	1	A 67	20

There are problematic predicates

Figure 6: Mark for problematic predicates

## 5. SUMMARY AND CONCLUSIONS

In this paper we have presented the tool created in the TAGFACT project, whose main objective is to create a tool to automatically annotate factuality in Spanish. This task has become especially relevant in the last few years in the field of Natural Language Processing. The multifaceted tool presented allows for corpus creation, management and annotation and has been used to create the TAGFACT corpus and the ‘Gold Standard.’

The corpus includes texts extracted from different Spanish newspapers, belonging to different political ideologies. The extraction was carried out in groups of three pieces of news, each from a different newspaper, covering the same event, which can provide information about how facts are accounted for in each of the papers. At present, the corpus contains 46,947 words in 176 pieces of news and the ‘Gold Standard’ consists of around 10,000 words.

With respect to corpus creation, the tool greatly facilitates inputting both the text and the metadata required for text identification. In addition, it presents a user-friendly interface to edit and manage the corpora created. Regarding the annotation of the corpus, the fact that it is linked to *Freeling* permits the automatic segmentation of texts into sentences and clauses. In this way, it can be used to annotate corpora at different levels, from whole texts to just words. The tool allows users to create their own labels, so it is possible to annotate linguistic information relevant to varied projects. Another relevant feature offered by the editor is the possibility of marking the voice of a predicate, that is, the narrator of the situation, or any other word that might trigger a decision for the various levels of annotation.

Currently, we are manually annotating the pieces of news of the ‘Gold Standard’ with regard to how events are presented with respect to author’s commitment. In TAGFACT, factual information is very rich and is inferred taking into account the four layers described in this paper that cover the different aspects taken into consideration in the project.

The two functionalities of the tool, corpus creation and corpus annotation, are versatile resources that can be freely used by any researcher working in Corpus Linguistics. The tool will be made available on the Internet under a *GNU General Public License*. In the future, we aim to complete the implementation of the system for the automatic annotation of factuality for Spanish.

#### REFERENCES

- Agerri, Rodrigo, Josu Bermúdez and German Rigau. 2014. Ixa pipeline: Efficient and ready to use multilingual NLP tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Reykjavik: European Language Resources Association, 3823–3828.
- Alonso, Laura, Irene Castellón, Hortènsia Curell, Ana Fernández-Montraveta, Sònia Oliver and Glòria Vázquez. 2018. Proyecto TAGFACT: Del texto al conocimiento. Factualidad y grados de certeza en español. *Procesamiento del Lenguaje Natural* 61: 151–154.
- Diab, Mona, Bori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, Vinodkumar and Weiwei Guo. 2009. Committed belief annotation and tagging. In Manfred Stede, Chu-Ren Huang, Nancy Ide and Adam Meyers eds. *Proceedings of the Third Linguistic Annotation Workshop*. Singapur: Association for Computational Linguistics, 68–73.
- Huang, Rongtao, Zou Bowei, Wang Hongling, Li Peifeng and Zhou Guodong. 2019. Event factuality detection in discourse. In Jie Tang, Min-Yen Kan, Dongyan

- Zhao, Sujian Li and Hongying Zan eds. *Natural Language Processing and Chinese Computing*. NLPCC 2019. Lecture Notes in Computer Science. Vol. 11839. Springer, Cham, 404–414.
- Krause, Thomas and Amir Zeldes. 2016. ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities* 31/1: 118–139.
- Lee, Kenton, Yoav Artzi, Yejin Choi and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In Lluís Màrquez, Chris Callison-Burch and Jian Su eds. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon: Association for Computational Linguistics, 1643–1648.
- Lloberes Marina, Irene Castellón, Lluís Padró. 2015. Suitability of ParTes test suite for parsing evaluation. *Proceedings of the 14<sup>th</sup> International Conference on Parsing Technologies*. Bilbao: Association for Computational Linguistics, 61–65.
- Marneffe, Marie-Catherine, Christopher D. Manning and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics* 38/2: 301–333.
- Matsuyoshi, Suguru, Megumi Eguchi, Chitose Sao, Koji Murakami, Kentaro Inui and Yuji Matsumoto. 2010. Annotating event mentions in text with modality, focus and source information. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner and Daniel Tapias eds. *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. Valetta: European Language Resources Association, 1456–1463.
- Minard, Anne-Lyse, Manuela Speranza and Tommaso Caselli. 2016. Event factuality annotation task (FactA). In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro and Rachele Sprugnoli eds. *Proceedings of the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*. Napoli: Open Edition Books, 32–39.
- Mullick, Ankan, Sourav Pal, Projjal Chanda, Arijit Panigrahy, Anurag Bharadwaj, Siddhant Singh and Tanmoy Dam. 2019. D-FJ: Deep neural network based factuality judgment. *TrueFact, Truth Discovery and Fact Checking: Theory and Practice workshop*.
- Narita, Kazuya, Junta Mizuno and Kentaro Inui. 2013. A lexicon-based investigation of research issues in Japanese factuality analysis. In Ruslan Mitkov and Jong C. Park eds. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Nagoya: Asian Federation of Natural Language Processing, 587–595.
- O'Donnell, Mick. 2008. The UAM CorpusTool: Software for corpus annotation and exploration. In Carmen Bretones, José Francisco Fernández, José Ramón Ibáñez, M. Elena García, M. Enriqueta Cortés, Sagrario Salaberri, M. Soledad Cruz, Nobel Perdu and Blasina Cantizano eds. *Applied Linguistics Now: Understanding Language and Mind*. Almería: Universidad de Almería, 1433–1447.
- Ogren, Philip V. 2006. Knowtator: A protégé plug-in for annotated corpus construction. In Alex Rudnicky, John Dowding and Natasa Milic-Frayling eds. *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Demonstrations*. New York: Association for Computational Linguistics, 273–275.
- Padró, Lluís and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet

- Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the Eight International Conference on Language Resources and Evaluation*. Istanbul: European Language Resources Association, 2473–2479.
- Prabhakaran, Vinodkumar, Tomas By, Julia Hirschberg, Owen Rambow, Samira Shaikh, Tomek Strzalkowski, Jennifer Tracey, Michael Arrigo, Rupayan Basu, Micah Clark, Adam Dalton, Mona Diab, Louise Guthrie, Anna Prokofieva, Stephanie Strassel, Gregory Werner, Yorick Wilks and Janyce Wiebe. 2015. A new dataset and evaluation for belief/factuality. *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*. Denver: Association for Computational Linguistics, 82–91.
- Saurí, Roser. 2008. *A Factuality Profiler for Eventualities in Text*. Massachusetts: Brandeis University dissertation.
- Saurí, Roser and James Pustejovsky. 2009. FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation* 43/3: 227–268.
- Soni, Sandeep, Tanushree Mitra, Eric Gilbert and Jacob Eisenstein. 2014. Modeling factuality judgments in social media text. In Kristina Toutanova and Hua Wu eds. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Volume 2: Short Papers*. Baltimore: Association for Computational Linguistics, 415–420.
- Soraa, Aitor, German Rigau, Jordi Porta, Jordi Atserias, Xavier Gómez Guinovart and Horacio Saggion. 2017. *Plataformas y Sistemas de Procesamiento Lingüístico de Alto Rendimiento*. Plan de impulso de las tecnologías del lenguaje: Ministerio de Energía Turismo y la Agenda Digital.
- Tonelli, Sara, Rachele Sprugnoli and Manuela Speranza. 2014. NewsReader guidelines for annotation at document level. In extension of deliverable D3. *Technical Report NWR-2014-2*. Trento.
- van Son, Chantal, Marieke van Erp, Antske Fokkens and Piek Vossen. 2014. Hope and fear: Interpreting perspectives by integrating sentiment and event factuality. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Reykjavik: European Language Resources Association, 26–31.
- Vázquez, Gloria and Ana Fernández-Montraveta. In press. Annotating factuality in the TAGFACT corpus. Comares.
- Velupillai, Sumithra. 2011. Automatic classification of factuality levels. A case study on Swedish diagnoses and the impact of local context. In Anne Moen, Stig Kjaer Andersen, Jos Aarts and Petter Hurlen eds. *User Centred Networked Health Care Proceedings of the European Federation of Medical Informatics*. Amsterdam: IOS Press, 559–563.
- Weisser, Martin. 2016. DART – The dialogue annotation and research tool. *Corpus Linguistics and Linguistic Theory* 12/2: 355–388.
- Wonsever, Dina, Aiala Rosá and Marisa Malcuori. 2016. Factuality annotation and learning in Spanish texts. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asunción Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the Tenth Conference on Language Resources and Evaluation*. Portoroz: European Language Resources Association, 2076–2080.

*Corresponding author*

Ana Fernández-Montraveta  
Autonomous University of Barcelona  
Facultat de Filosofia i Lletres  
08193 Bellaterra  
Cerdanyola del Vallès · Barcelona  
Spain  
e-mail: Ana.Fernandez@uab.cat

received: January 2020  
accepted: April 2020

# *The Primary Education Learners' English Corpus (PELEC): Design and compilation*

Zeltia Blanco-Suárez – Francisco Gallardo-del-Puerto – Evelyn Gandón-Chapela  
University of Cantabria / Spain

**Abstract** – This paper describes the process of design and compilation of the *Primary Education Learners' English Corpus* (PELEC), a learner corpus which includes written (14,577 words) and spoken materials (47,032 words) from Primary Education learners in the Autonomous Community of Cantabria. It is composed of data from a total of 252 students in the fourth and sixth grade of Primary Education (aged 9–10 and 11–12, respectively) who were studying in five different state schools which followed either a Content and Language Integrated Learning (CLIL) or an English as a Foreign Language (EFL) approach.

**Keywords** – PELEC; learner corpora; Content and Language Integrated Learning (CLIL); English as a Foreign Language (EFL); young learners; Primary Education

## 1. INTRODUCTION<sup>1</sup>

Over the past three decades, the role of computer learner corpora, i.e. “systematic computerized collections of texts produced by language learners” (Nesselhauf 2004: 125), has become of paramount importance in the field of Second Language Acquisition and Teaching. Thanks to the technological advances currently available, it is now possible to store and tag language learners’ oral and written productions in electronic format, as well as to retrieve and analyse them automatically. This fact opened up a wide range of research possibilities given that scholars could access learners’ data much more quickly and easily, instead of having to store and consult them in “shoe boxes” (Díez-Bedmar 2009: 920). It is in this light that numerous learner corpora have emerged, especially in Europe (see, among many others, the *International Corpus of Learner English* (ICLE),

---

<sup>1</sup> The authors would like to thank Dr. Julia Williams Camus for her invaluable help in the design and compilation of PELEC. We would also like to thank two research assistants (Míriam Fernández Arenal and Silvia Mendiguchía Pérez) for their participation in the data gathering procedure. For generous financial support, we are grateful to the Vice-rectorate for Research and Transfer of Knowledge from the University of Cantabria (grant ref. UC2016-GRE-10). Special thanks go to Alberto San Emeterio Bolado for the storytelling panels. We would also like to thank two anonymous reviewers for their helpful comments and suggestions.



initiated by Granger (Catholic University of Louvain) in 1990 and first published in 2002; see Granger *et al.* 2002). In the particular context of Spain, several English learner corpora have been compiled with data from primary, secondary and university students (see Section 2 for more details).

This paper presents the characteristics of *The Primary Education Learners' English Corpus* (henceforth PELEC), which was compiled in 2018 by a team of researchers at the University of Cantabria with the aim of gathering data from primary students in the Autonomous Community of Cantabria. This corpus includes both written and spoken materials from Primary Education learners of English as a second language and totals 61,609 words. At the time of data collection, the participants were enrolled in the fourth and sixth grade of Primary Education at five different state schools in Cantabria, which offered either traditional English as a Foreign Language (EFL) or Content and Language Integrated Learning (CLIL) programmes. All the schools selected were located in the outskirts of Santander and their choice was motivated by the rather homogeneous socio-economic status of the families in these institutions, in contrast to urban schools, which provided a sharper contrast in that regard.

As is well-known in the field of learner corpora, collecting data from young learners from this level of education poses a great number of additional challenges because of the technical difficulties involved. To mention but a few, data collection from this population involved obtaining ethical consent from the young learners' parents or legal guardians and the compliance with the current data protection regulations, especially so because the students were both audio recorded and videotaped when performing the oral tasks. However, even if the procedures of data collection and processing were rather complex and time-consuming, they were worth the effort in that they offer the possibility of analysing both the written and oral data from the same participant. This fact alone is innovative and opens a number of research avenues which have been relatively unexplored so far. Moreover, given that PELEC contains data from CLIL and non-CLIL state schools in Spain, the results gathered from its analysis may help to advance not only in the field of Second Language Acquisition research, but also in the realms of language teaching, language planning and language policies.

In the remainder of this paper, we will set forth the process of compilation and main characteristics of PELEC, a learner corpus featuring spoken and written data from L1-Spanish young learners of English. To this end, Section 2 first provides an initial

overview of the extant learner corpora which are most relevant to PELEC. Section 3 is concerned with the process of design and compilation of the corpus, including the participants of the study and the types of data collected. Finally, in section 4 the potential applications of PELEC and issues for further research are presented.

## 2. LEARNER CORPORA: AN OVERVIEW

Learner corpora offer manifold research possibilities, although they have been mainly used for two purposes, namely to gain insights into second or foreign language acquisition and to create tailor-made pedagogical materials based on the students' most frequent errors (Granger 2008: 259). This section will offer a brief account of the learner corpora available with data from L1-Spanish learners of English. As will be shown, these learner corpora have targeted different levels of education, ranging from primary through to secondary and university levels. Some of them contain both oral and written data, whereas some others have focused only on written data. In what follows, a classification of the levels and types of data included in the different corpora is offered. Firstly, we will present the characteristics of the corpora based on data obtained from primary and secondary students, to then focus on those corpora that only contain data from university level pupils.

Among the corpora that compile data from primary and secondary students, one can find *The Barcelona Age Factor (BAF) corpus*, *The Barcelona English Language Corpus (BELC)* and the corpora used by the research teams *Research in English Applied Linguistics (REAL)* and *Language and Speech Laboratory (LASLAB)*, based at the University of Barcelona and the University of the Basque Country, respectively.<sup>2</sup> In particular, the BAF corpus and the BELC corpus gathered both oral (oral narratives, role plays, oral interviews) and written data (written compositions) to analyse the effect played by the age at which students started to learn English (2,063 participants in total). The BELC corpus, stemming from the BAF corpus, allows to track subjects longitudinally over a period of seven years (see Muñoz 2006). The participants in these corpora were all bilingual in Spanish and Catalan and studied in state schools in Catalonia, differing only in the starting age of English instruction (8, 11, 14 and 18).

---

<sup>2</sup> For further information on the BAF and BELC corpora, see <https://slabank.talkbank.org/access/English/BELC.html>.

The age factor was also the focus of a number of projects stemming from REAL and LASLAB. This team of researchers gathered data from bilingual (Basque and Spanish) students enrolled at various primary and secondary schools in this region, which offered either CLIL programmes or traditional EFL classrooms. The participants, who had started learning English at different ages (4, 8 and 11), were asked to do several written and oral tasks, as well as different reading and listening activities, so as to measure their overall language proficiency. This variety of data enabled researchers to shed light on a wide range of aspects relevant to the acquisition and learning of English, including the maturational effects in L3 English, the learning context and the type of task performed (see, among many others, García-Mayo and García-Lecumberri 2003; Lasagabaster and Doiz 2003; Gallardo-del-Puerto and Gómez-Lacabex 2013, 2017; García-Mayo and Imaz-Agirre 2019).

Several learner corpora have also sourced from secondary and university students. This is the case, for instance, of the *Santiago University Learner of English Corpus* (SULEC), the *Universidad Autónoma de Madrid* (UAM) *Corpus de Interlenguas Escritas* and the *Corpus of English as a Foreign Language* (COREFL).

The first of these resources, SULEC, is a monitor corpus compiled at the University of Santiago de Compostela under the direction of Ignacio Palacios-Martínez (see Palacios-Martínez 2005). The project intended to compile 1,000,000 words of oral and written English by primary, secondary and university students with different proficiency levels. At present, the corpus comprises around 500,000 words<sup>3</sup> and it only contains data from secondary (first and second year *Bachillerato*) and university students from the degree in English Philology with intermediate and advanced levels of proficiency. The written data were collected in the form of 500-word compositions, whereas the oral component was collected through instruments such as oral presentations, oral exams and personal interviews.

The second corpus which also includes data from secondary and university students, the so-called *UAM Corpus de Interlenguas Escritas*, is divided into three subcomponents. One of them contains 210 essays written by secondary school pupils during class time. This subcomponent includes 174 texts that were written by 87 students from the first, second and third years of *Bachillerato* and the former pre-university *Curso*

---

<sup>3</sup> Palacios-Martínez, January 2020, personal communication.

*de Orientación Universitaria* (COU) before and after a pedagogical intervention. The remaining 36 texts belong to students from the same levels of education but do not have a pre- or post-task counterpart (Barrio-Luis 2005: 64–65; Díez-Bedmar 2009: 925). The main results of this intervention programme were published in a book by Martín-Úriz and Whittaker (2005). Another subcomponent is formed by a collection of essays written by 119 pre-university students from several high-schools who wrote about three composition topics and answered a cloze test (Martín-Úriz and Whittaker 2005; Díez-Bedmar 2009: 925). The third subcomponent of this corpus is composed by essays on the same topic as the first subcomponent, which were written by students in their first year of the degree in English Philology at the Autonomous University of Madrid (Díez-Bedmar 2009). The process of error-tagging of this corpus and the toolbar used to tag errors and retrieve them have been described in Barrio-Luis (2005).

The COREFL, on its part, is a corpus that is currently being compiled at the Universities of Granada and Bremen. It contains L2 English written and spoken data from L1 Spanish and L1 German learners at secondary and university levels with different proficiency levels (A1-C2 in the Common European Framework of Reference for Languages) and ages (from 12 years onwards). As of September 2019, according to Díaz-Negrillo *et al.* (2019), the corpus contained approximately 1,612 texts (189 oral and 1,423 written texts) sampling four different types of narrative tasks. In the future, this corpus will also feature two control corpora of the learners' L1, that is, Spanish (including Peninsular and Latin American varieties) and German (under compilation). Interestingly, the L1 Spanish-L2 English subcomponent includes data from learners in different types of instructional contexts: secondary school bilingual programmes (CLIL) vs. mainstream EFL classrooms, on the one hand, and university English as a Medium of Instruction (EMI) learners vs. university Spanish as a Medium of Instruction (SMI) learners, on the other.

By contrast, some other corpora restricted their samples to university students, as is the case of the *Madrid Corpus* (MAD), the *Written Corpus of Learner English* (WriCLE), the *English Written Interlanguage* (ENWIL) and the *Non-native Spanish Corpus of English* (NOSE).

MAD was compiled at the Complutense University of Madrid by the SPAINWRITE research group and is subdivided into three components. The first one is composed of a collection of argumentative essays written by over 200 students of English

as a foreign language from the degree in English Philology in their first and fourth years. The second subcomponent includes argumentative essays by the same students in their native language, and finally, the third subcomponent consists of a control corpus of essays written by third-year American students of the degree in Spanish Philology that formed part of the Middlebury Programme in Madrid (Díez-Bedmar 2009: 923).

The *Written Corpus of Learner English* (WriCLE; Rollinson and Mendikoetxea 2010), on its part, includes two subcomponents. The first, WriCLEformal, contains a set of 752 essays written by Spanish university students (from all levels of proficiency) in their first and third years of the degree in English Studies at Autonomous University of Madrid (around 750,000 words in XML format). The second subcomponent, WriCLEinf, is the informal, non-academic counterpart of the WriCLEformal, featuring texts from blogs, emails, autobiographical pieces of writing, narratives, descriptions, poems, among many others, amounting to 1,140 texts and totalling around 8,000 words.<sup>4</sup>

Another learner corpus drawing upon collections of university students' essays is the *English Written Interlanguage* (ENWIL) corpus, created in 1997. ENWIL includes essays written by first-year students of English Philology at the University of Alcalá de Henares (Valero-Garcés *et al.* 2000). In line with the *UAM Corpus de Interlenguas Escritas*, it has also been error-tagged and the ultimate aim was to create more tailor-made materials targeting the students' needs, which resulted in a resource book for Spanish learners of English on how to write successfully in the realm of academic writing (Valero-Garcés *et al.* 2003).

Likewise, the *Non-native Spanish Corpus of English* (NOSE), of about 300,000 words from 1,000 samples of 250-300 words, is a collection of descriptive and argumentative essays written by approximately 500 Spanish students of English at the universities of Granada and Jaén. The corpus is also error-tagged and is available with a corpus tool, which allows to search for a number of variables, including the informants' profiles, topics and text types (Díaz-Negrillo 2012). This has not only permitted researchers to assess and diagnose the written competence of the learners, but also to "propose remedial work to counteract students' difficulties" (Díaz-Negrillo 2012: 43).

In addition to the aforementioned L1-Spanish English learner corpora, there also exist a number of learner corpora with different mother tongue backgrounds. Among such

---

<sup>4</sup> For more detailed information on this corpus, visit <http://wricle.learnercorpora.com/>.

initiatives we find the *International Corpus of Crosslinguistic Interlanguage* (ICCI), which constitutes an international joint project initiated by Prof. Yukio Tono (Tokyo University of Foreign Studies) in 2007 (Hong 2012: 47). This corpus contains 6,700 transcripts of argumentative and descriptive essays written by students from grades 3 to 12, i.e. from primary and secondary education levels. This over 500,000-word corpus represents 6,700 subjects from seven different countries (Austria, China, Hong Kong, Israel, Poland, Spain and Taiwan), thirty-five mother tongues and different proficiency levels (see Hong 2012: 50–51 for more details). Likewise, Sylviane Granger launched another project in Europe, the *International Corpus of Learner English* (ICLE), whose first version was published on CD-ROM in 2002 and which is available since 2009 as an expanded version, ICLEv2 (see Granger *et al.* 2002). ICLE comprises argumentative essays by intermediate to advanced learners of English from languages as diverse as Bulgarian, Chinese, Dutch, Finnish or Turkish, totalling about 3.7 million words. The compilation of the Spanish subcomponent of ICLE, known as SPICLE, was undertaken by the compilers of MAD (Martínez-Osés and Neff-van Aertselaer 2001). ICLE has a spoken counterpart, the *Louvain International Database of Spoken English Interlanguage* (LINDSEI), which includes interviews with EFL learners with different mother tongues, for a total of about 100,000 words (Gilquin *et al.* 2010). Following the same principles and guidelines in ICLE and LINDSEI, the Centre for English Corpus Linguistics at the Catholic University of Louvain additionally created the *Longitudinal Database of Learner English* (LONGDALE), which, unlike the spoken and written corpora, is not exclusively focused on interviews or argumentative essays, but contains a wide variety of data, including grammaticality judgement tests.<sup>5</sup> At a much larger scale than ICLE and LINDSEI, Cambridge University Press and Longman have devised two commercial mega-corpora which are constantly being expanded: the *Longman Learners' Corpus* and the *Cambridge Learner Corpus* (CLC). These corpora contain over 10 million words and represent countless mother tongue backgrounds (Granger 2008: 261).

Having provided a broad overview of the main corpora of L1-Spanish learners of English, in the following section we will offer a detailed description of the design of PELEC and its compilation, including data regarding the participants, the different

---

<sup>5</sup> For more information on LONGDALE, see <https://uclouvain.be/en/research-institutes/ilc/cecl/longdale.html>.

questionnaires and tests used to gather the data, as well as an overview of the written and spoken components of the corpus.

### 3. CORPUS DESIGN AND COMPILATION

#### 3.1. *Participants*

The corpus gathers data from a total of 252 students from the fourth and sixth grade of Primary Education (aged 9-10 and 11-12, respectively) in five different state schools in Cantabria. In compliance with current data protection regulations, the participants' parents or legal guardians were asked to sign a consent provided by the University of Cantabria before the data compilation process started in each of the schools. Three of these schools offered a CLIL approach, while in the remaining two learners received regular EFL courses. Therefore, in the non-CLIL groups students were exposed to three weekly hours of English, as required in the curriculum. In turn, in the CLIL groups learners benefited from two extra CLIL hours of English in addition to the compulsory ones, in subjects such as Natural Sciences, Physical Education, Arts and Crafts or Music. Table 1 shows the distribution of the students in the corpus according to the type of approach, gender, grade and number of hours of instruction.

	Students	Gender	English exposure	
			Grade 4	Grade 6
<b>CLIL</b>	124	F: 46.77% (n=58)	EFL 361h	EFL 617 h
		M: 53.23% (n=66)	CLIL 307 h	CLIL 462 h
<b>Non-CLIL</b>	128	F: 53.12% (n=68)	EFL 361 h	EFL 617 h
		M: 46.88% (n=60)		

Table 1: Description of participants in PELEC

#### 3.2. *Types of materials*

##### 3.2.1. Questionnaires and tests

In order to obtain a more comprehensive picture of their English learning profiles, all the students were asked to complete an initial questionnaire in Spanish consisting of biographical information (parents' or legal guardians' occupations, L2 learning onset time, extramural exposure, etc.) and of the compensatory strategies pursued when attempting to overcome a communicative gap, such as L1 use, appeal for assistance, paraphrasing, etc. (see, among others, O'Malley and Chamot 1990; Gallardo-del-Puerto

and Gómez-Lacabex 2017). Together with these initial questions, they were also asked to answer a motivation questionnaire in Spanish, which was based on Gardner's (1985) Attitude/Motivation Test Battery (AMTB) and adapted to this type of learners, in line with previous motivation studies for young apprentices (Kiss and Nikolov 2005; Carreira 2006; Cid *et al.* 2009; Lasagabaster and Sierra 2009; Fernández-Fontecha 2014, 2015). This test comprised a total of 34 items measuring factors such as their reported effort and self-efficacy, their willingness to integrate in the target language community, their anxiety levels, the degree of parental support, as well as their intrinsic and extrinsic motivation. All the statements in these questionnaires had to be marked on a 5-point Likert scale, from the lowest (*I do not agree at all* 😞😞) to the highest degree (*I completely agree* 😊😊). The results from the motivation survey revealed that the motivation profiles of the CLIL and non-CLIL students were rather similar (see Gallardo-del-Puerto and Blanco-Suárez forthcoming): both groups exhibited particularly high levels of extrinsic motivation (the external factors to learn the language), which should not be surprising given the foreign, non-naturalistic learning setting. Interestingly, CLIL students reported being more encouraged by their parents or families to learn English than their EFL counterparts. As for the effect of gender, no differences were found in the motivation scores of male vs. female young learners in the CLIL group. Nonetheless, in the non-CLIL context girls outperformed boys in the overall and intrinsic motivation. This motivation questionnaire, together with the one on background information and the one on compensatory strategies, were completed during a 50-minute session.

In addition to the aforementioned questionnaires, the students' competence in English was examined by means of a series of language tests (see Table 2). Thus, they had to do a listening, reading and a use of English test, for which purposes we drew on materials from the Cambridge English A1 Movers and A2 Flyers tests. The listening comprehension test consisted of two multiple choice exercises (with five items each) and the reading comprehension included three short texts which described one student each and their daily routines. In total, they had to answer ten multiple choice questions related to these characters and their lives. For the use of English test, our young learners had to complete a cloze test with ten gaps in two emails which were exchanged between a Spanish and a Chinese student. The blanks related to grammatical contents such as the article use, the third person present tense inflection *-s* and the use of pronouns or

prepositions, in accordance with the curriculum for those educational levels. These language tests were conducted on two separate days during a 50-minute session.

	<b>Use of English (max=10)</b>	<b>Listening (max=10)</b>	<b>Reading (max=10)</b>
Non-CLIL Grade 4	4.24	5.39	3.36
CLIL Grade 4	4.87	5.25	3.49
Non-CLIL Grade 6	6.55	7.23	4.76
CLIL Grade 6	6.69	7.02	5.33
<b>Non-CLIL all</b>	<b>5.42</b>	<b>6.29</b>	<b>4.07</b>
<b>CLIL all</b>	<b>5.52</b>	<b>5.84</b>	<b>4.13</b>

Table 2: Mean scores in the language tests

As can be seen in Table 2, the mean scores obtained by CLIL and non-CLIL learners were rather similar, more differences being observed when grade 4 vs. grade 6 students are compared, either in the CLIL or the non-CLIL samples.

### 3.2.2. Written component

The written component of PELEC comprises 246 compositions of L1-Spanish learners of English, totalling 14,577 words (6,398 and 8,179 words from fourth and sixth grade, respectively), as shown in Table 3. The average length of these writings is 58.08 words, ranging from a minimum of 4 to a maximum of 191 words, and the standard deviation is +/-33.401.

	<b>4<sup>th</sup> grade</b>	<b>6<sup>th</sup> grade</b>	<b>Both grades</b>
	No. of words	No. of words	No. of words
Non-CLIL	2,870	4,629	7,499
CLIL	3,528	3,550	7,078
<b>Total</b>	<b>6,398</b>	<b>8,179</b>	<b>14,577</b>

Table 3: Written component of PELEC

For this part, students were asked to write a short letter to an English pen friend, Tom, telling him about their favourite things and what they do on a normal day. They had to complete this task in approximately 20 minutes. Moreover, they were asked to write the same letter in Spanish some weeks later, which would additionally allow us to verify their level of written competence in their first language.

All the writings were scanned and later transcribed and saved in separate .txt files, so that the texts could be prepared for subsequent analysis with a concordance software and with the CLAN tool from *TalkBank* software.<sup>6</sup> Each file contains the body of the text

<sup>6</sup> For more information on CLAN and *TalkBank*, see <https://childes.talkbank.org/>.

(a student writing) with its corresponding header, which includes the following metadata: filename, school name, grade and group, date of collection, as well as the student's name(s) and surnames. The filename additionally identifies the type of task, school and year. Data regarding the students' language competence in the other language tests (reading, listening and use of English) and questionnaires were stored on a separate Excel spreadsheet with all the participants. The corpus has thus far not been tagged for part of speech (POS) and no linguistic annotation has yet been added to mark any relevant lexical or morphosyntactic features. Nonetheless, corpus annotation would be possible in the .txt files with XML-language, in compliance with the TEI guidelines, and in CLAN. The following excerpt from one of the students' written compositions serves to illustrate the written component of PELEC:

- (1) \*CHI: at eight o'clock I go at the kitchen to have lunch.  
 \*CHI: at half past eight I'm wash my teeths and I dress up.  
 \*CHI: I take the bag and I go to school.

### 3.2.3. Oral component

As shown in Table 4, the spoken component of PELEC includes a total of 47,032 words, 24,863 words from fourth-grade student spoken productions and 22,169 words of spoken materials from sixth-grade students. The average length of students' oral productions is 181.05 words,<sup>7</sup> with a standard deviation of +/-89.457, a minimum of 11 words and a maximum of 572. The oral corpus contains a total of 7,771 utterances with a mean of 32.66 utterances per student, the minimum and maximum being 4 and 82 utterances, respectively. The average length of individual utterances is of 6.46 words, the standard deviation being +/-3.981 and the minimum and maximum ranging from 1 to 39.8 words per utterance.

	<b>4<sup>th</sup> grade</b>	<b>6<sup>th</sup> grade</b>	<b>Both grades</b>
	No. of words	No. of words	No. of words
Non-CLIL	11,092	13,435	24,527
CLIL	13,771	8,734	22,505
<b>Total</b>	<b>24,863</b>	<b>22,169</b>	<b>47,032</b>

Table 4: Spoken component of PELEC

<sup>7</sup> Given that PELEC consists of three different tasks and that the data for each of them were collected on various days, the number of students who accomplished each task may differ.

For the spoken data of PELEC, our young learners were taped and video recorded performing two separate tasks. In the first one, they were given a set of pictures and asked to tell the story individually (see Figure 1). On average, the students completed this task in five minutes. Additionally, they were asked to tell the story in Spanish some weeks later, which would allow us to compare these children's spoken behaviour in the L1 and the L2.



Figure 1: Speaking task I: Storytelling

The drawings show a friendly dog which seems to be lost on a rainy day. Luckily, the protagonist of the story, a young boy holding an umbrella, runs into the dog and they walk home together. There they are welcomed by the boy's parents and the dog becomes a new member of the family. The students were asked to perform this task following their own resources, without any help. The researcher who was recording the spoken productions did not intervene at any moment, unless required by the child in cases of appeal for assistance or in cases in which the students had a mental block and were not able to continue with the activity. Example (2) provides an extract of a student's oral production in this task.

- (2) \*CHI: the person with the umbrella eeheh (3.) takes the dog.  
 \*CHI: a:nd (3.) travel with him to her house.  
 \*CHI: and they take dog for us.

The second was a spot-the-differences task which was done in pairs, in line with an exercise in the Cambridge Young Learners English Test (Movers, A1 level). By asking each other questions, dyads had to collaborate to find five differences in their respective photos (see Figure 2), which took them approximately ten minutes. Since there was a barrier between them, they could not see each other's photos, so they were forced to rely

on the linguistic resources at hand to discover the differences. Both oral production activities were recorded and later transcribed by two coordinated research assistants. As in the case of the storytelling, the researcher(s) present during the recording session did not take part unless specifically requested by the students or in those cases in which they became too nervous and were unable to follow.

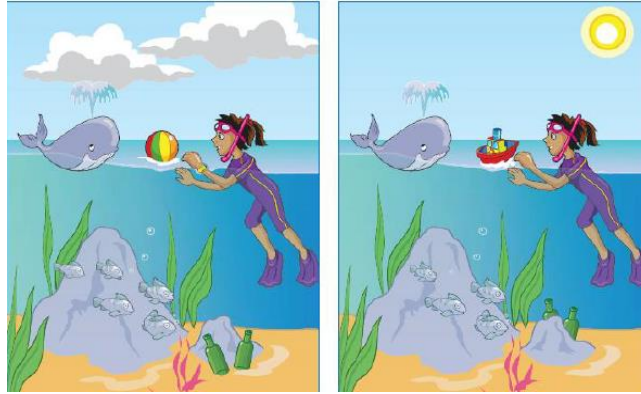


Figure 2: Speaking task II: Spot the differences<sup>8</sup>

For the data transcription conventions, we issued some guidelines based on other spoken corpora, including LINDSEI, the Lancaster Corpus and VOICE, as well as on CLAN and *TalkBank*. Thus, for each transcription we marked the participants' names, including that of the researcher who was present during the recording session, the filename, the transcriber's name, the date of the recording and then the beginning and end of each of the interventions. Moreover, we marked each speaker's turn, any overlaps in the different turns, pauses or lengthening of an utterance, laughter and non-verbal elements (e.g. coughing or body language), L1 use and the use of fillers. Example (3) illustrates part of the interaction between two students in this task:

- (3) \*CH1: aaam (7.) in your pictures the bottles is green?  
 \*CH2: the eeoh yes.  
 \*CH2: eeem eeoh i:n your picture the girl wear a ponytail?  
 \*CH1: yes.

#### 4. CORPUS APPLICATIONS AND ISSUES FOR FURTHER RESEARCH

PELEC opens up a wide range of research possibilities. Firstly, the potentialities of a corpus of this kind in the field of language acquisition research are immense, since it

<sup>8</sup> Taken from the online A1 Movers sample test at <https://www.cambridgeenglish.org/Images/young-learners-sample-papers-2018-vol1.pdf>.

would allow numerous studies on phonological, lexical, morphosyntactic and discourse aspects, thereby detecting the most problematic areas for learners in both CLIL and non-CLIL learning contexts. In this regard, the first study derived from the investigation of the present corpus analysed verb number agreement errors and null subjects (Fernández-Pena and Gallardo-del-Puerto 2019) in Primary Education Grade 6 schoolchildren. This investigation did not find striking differences between the CLIL and non-CLIL samples examined concerning these two aspects of English grammar. Both groups omitted expletive subject pronouns (*\*is raining* vs. *it is raining*) to a considerably larger extent than referential subject pronouns (*\*is a dog* vs. *it is a dog*), the latter being dropped minimally. Similarly, they both produced omission errors more frequently in affixal (*\*the boy sleep with the dog* vs. *the boy sleeps with the dog*) than in suppletive inflection (*\*the boy sleeping with the dog* vs. *the boy is sleeping with the dog*). However, non-CLIL learners omitted auxiliary *be* (*\*the boy sleeping with the dog* vs. *the boy is sleeping with the dog*) more frequently than copula *be* (*\*the dog in a bedroom* vs. *the dog is in a bedroom*), which, together with their greater use of null expletive subjects and of placeholders (*\*the boy is sleep with the dog* vs. *the boy sleeps with the dog*), was indicative of an earlier stage of acquisition. Conversely, although the presence of commission errors (*\*they goes to bed* vs. *they go to bed*) was minimal in the data, CLIL learners' rate of incorrect inflection supply in copula *be* contexts (*\*your eyes is brown* vs. *your eyes are brown*) was surprisingly higher than that of non-CLIL learners. In addition to the analyses reported by Fernández-Pena and Gallardo-del-Puerto (2019), PELEC would permit the analysis of, for instance, measures of amount of production (type and token ratios), density of production (total number of tokens per utterance, etc.) and compensatory strategies, such as appeals for assistance or L1 use, in line with the studies by Gallardo-del-Puerto and Gómez-Lacabex (2013, 2017).

Secondly, cross-linguistic studies would also be possible with PELEC, given that, as was the case with MAD and COREFL, PELEC records L1-data in the writing and in the storytelling tasks. In addition, the analysis of the spot-the-differences task in English would allow a comparison with the results from previous investigations on collaborative interaction such as García-Mayo and Imaz-Agirre (2019) and Martínez-Adrián (2020). The former discovered that young CLIL learners' occurrence of language-related episodes depended on the type of task, whereas the latter found that older schoolchildren

resorted to previously known languages more frequently than younger ones to keep the flow of interactive speech.

Thirdly, the analyses derived from the students' L2-productions in this corpus would be highly beneficial to educators, as another possible application would be the creation of more targeted and tailor-made materials based on the most recurrent errors, in line with the aims of some of the corpora mentioned in Section 2. Importantly, the analysis of the output of the CLIL vs. non-CLIL learning contexts would also enable the contribution to the realms of language planning and language policies in Spain in the long run.

Finally, although the corpus presents several limitations in terms of the student sample and its size, it could be expanded by including representation from all grades in Primary Education and additional schools in Cantabria, both from CLIL and non-CLIL approaches. Furthermore, the process of data collection and transcription could be extended to other educational levels such as Secondary and Tertiary Education. This would of course allow us to obtain a more comprehensive picture of the productive skills of the English learners in this northern region in Spain.

## REFERENCES

- Barrio-Luis, María. 2005. Diseño del corpus de interlenguas de textos escritos en inglés lengua extranjera. In Ana Martín-Úriz and Rachel Whittaker eds. *La Composición como Comunicación: Una Experiencia en las Aulas de Lengua Inglesa en Bachillerato*. Madrid: Ediciones de la Universidad Autónoma de Madrid, 61–75.
- Carreira, Junko Matsuzaki. 2006. Motivation for learning English as a foreign language in Japanese elementary schools. *JALT Journal* 28/2: 135–158.
- Cid, Eva, Gisela Grañena and Elsa Tragant. 2009. Constructing and validating the foreign language attitudes and goals survey (FLAGS). *System* 37/3: 496–513.
- Díaz-Negrillo, Ana. 2012. Learner corpora: The case of the NOSE corpus. *Systemics, Cybernetics and Informatics* 10/1: 42–47.
- Díaz-Negrillo, Ana, Cristóbal Lozano and Marcus Callies. 2019. Introducing the Corpus of English as a Foreign Language (COREFL): A bimodal, multi-task corpus for SLA research. Paper presented at the 5<sup>th</sup> Learner Corpus Conference (12–14 September 2019). University of Warsaw.
- Díez-Bedmar, María Belén. 2009. Written learner corpora by Spanish students of English: An overview. In Pascual Cantos-Gómez and Aquilino Sánchez-Pérez eds. *A Survey of Corpus-based Research: Proceedings of the AELINCO Conference*. Murcia: Asociación Española de Lingüística de Corpus, 920–933.
- Fernández-Fontecha, Almudena. 2014. Motivation and gender effect in receptive vocabulary learning: An exploratory analysis in CLIL Primary Education. *Latin American Journal of Content and Language Integrated Learning* 7/2: 27–49.

- Fernández-Fontecha, Almudena. 2015. Motivation and vocabulary breadth in CLIL and EFL contexts: Different age, same time of exposure. *Complutense Journal of English Studies* 23: 79–96.
- Fernández-Pena, Yolanda and Francisco Gallardo-del-Puerto. 2019. Number agreement errors and subject omission in CLIL vs. non-CLIL learners of English in Primary Education. Paper presented at the *43rd Conference of the Spanish Association of Anglo-American Studies* (13–15 November 2019). University of Alicante.
- Gallardo-del-Puerto, Francisco and Zeltia Blanco-Suárez (forthcoming). Foreign Language Motivation in Primary Education students: The effects of additional CLIL and gender. *Journal of Immersion and Content-based Language Education*.
- Gallardo-del-Puerto, Francisco and Esther Gómez-Lacabex. 2013. The impact of additional CLIL exposure on oral English production. *Journal of English Studies* 11/1:113–131.
- Gallardo-del-Puerto, Francisco and Esther Gómez-Lacabex. 2017. Oral production outcomes in CLIL: An attempt to manage amount of exposure. *European Journal of Applied Linguistics* 5/1: 31–54.
- García-Mayo, María Pilar and María Luisa García-Lecumberri eds. 2003. *Age and the Acquisition of English as a Foreign Language*. Clevedon: Multilingual Matters.
- García-Mayo, María del Pilar and Ainara Imaz-Agirre. 2019. Task modality and pair formation method: Their impact on patterns of interaction and LREs among EFL primary school children. *System* 80: 165–175.
- Gardner, Robert C. 1985. *Social Psychology and Second Language Learning: The Role of Attitudes and Motivation*. London: Edward Arnold.
- Granger, Sylviane. 2008. Learner corpora. In Anke Lüdeling and Merja Kytö eds. *Corpus Linguistics: An International Handbook* (Volume 1). Berlin: Mouton de Gruyter, 259–275.
- Gilquin, Gaëtanalle, Sylvie De Cock and Sylviane Granger eds. 2010. *LINDSEI: Louvain International Database of Spoken English Interlanguage*. Louvain: UCL Presses.
- Granger, Sylviane, Estelle Dagneux and Fanny Meunier. 2002. *The International Corpus of Learner English*. Louvain: Université Catholique de Louvain.
- Hong, Huaqing. 2012. Compilation and exploration of ICCI corpus for learner language research. In Yukio Tono, Yuji Kawaguchi and Makoto Minegishi eds. *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*. Amsterdam: John Benjamins, 47–62.
- Kiss, Csilla and Marianne Nikolov. 2005. Developing, piloting and validating an instrument to measure young learners' aptitude. *Language Learning* 55/1: 99–150.
- Lasagabaster, David and Aintzane Doiz. 2003. Maturation constraints on foreign-language written production. In María Pilar García-Mayo and María Luisa García-Lecumberri eds. *Age and the Acquisition of English as a Foreign Language*. Clevedon: Multilingual Matters, 136–160.
- Lasagabaster, David and Juan Manuel Sierra. 2009. Language attitudes in CLIL and traditional EFL classes. *International CLIL Research Journal* 1/2: 4–17.
- Martín-Úriz, Ana María and Rachel Whittaker eds. 2005. *La Composición como Comunicación: Una Experiencia en las Aulas de Lengua Inglesa en Bachillerato*. Madrid: Ediciones de la Universidad Autónoma de Madrid.
- Martínez-Adrián, María. 2020. The use of previously known languages and target language English during task-based interaction: A pseudolongitudinal study of primary-school CLIL learners. *EuroAmerican Journal of Applied Linguistics and Languages* 7/1: 59–77.

- Martínez-Osés, Francisco and Joanne Neff-van Aertselaer. 2001. Corpus analysis of prepositional patterns and non-native university writing. In Carme Muñoz-Lahoz, María Luz Celaya-Villanueva, Marta Fernández-Villanueva, Teresa Navés and Oliver Struck eds. *Trabajos en Lingüística Aplicada*. Barcelona: Univerbook, 139–147.
- Muñoz, Carmen ed. 2006. *Age and the Rate of Foreign Language Learning*. Clevedon: Multilingual Matters.
- Nesselhauf, Nadja. 2004. Learner corpora: Learner corpora and their potential for language teaching. In John McH. Sinclair ed. *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins, 125–152.
- O'Malley, J. Michael and Anna Uhl Chamot. 1990. *Learning Strategies in Second Language Acquisition*. Cambridge: Cambridge University Press.
- Palacios-Martínez, Ignacio. 2005. Las nuevas tecnologías y la investigación en el campo de la adquisición de segundas lenguas. In Mario Cal-Varela, Paloma Núñez-Pertejo and Ignacio Palacios-Martínez eds. *Nuevas Tecnologías en Lingüística, Traducción y Enseñanza de Lenguas*. Santiago de Compostela: Servizo de Publicacións da Universidade de Santiago de Compostela, 203–224.
- Rollinson, Paul and Amaya Mendikoetxea. 2010. Learner corpora and second language acquisition: Introducing WriCLE. In Jorge Luis Bueno-Alonso, Dolores González-Álvarez, Úrsula Kirsten-Torrado, Ana Elina Martínez-Insua, Javier Pérez-Guerra, Esperanza Rama-Martínez and Rosalía Rodríguez-Vázquez eds. *Analizar Datos > Describir variación/Analysing Data > Describing Variation*. Vigo: Servizo de Publicacións da Universidade de Vigo, 1–12.
- Valero-Garcés, Carmen, Guzmán Mancho-Barés, Carmen Flys-Junquera and Esperanza Cerdá-Redondo. 2000. Evolución de la interlengua y análisis de textos: *ENWIL* y el análisis de errores en la expresión escrita en EFL. In Francisco José Ruiz-de-Mendoza-Ibáñez, Lorena Pérez-Hernández, Mercedes Fornés-Guardia and Juan Manuel Molina-Valero eds. *Panorama Actual de la Lingüística Aplicada: Conocimiento, Pensamiento y Uso del Lenguaje. Vol. 3: Adquisición y Aprendizaje de Lenguas. Diseño Curricular. Lengua con Fines Específicos*. Logroño: Mogar Linotype, 1840–1860.
- Valero-Garcés, Carmen, Guzmán Mancho-Barés, Carmen Flys-Junquera and Esperanza Cerdá-Redondo. 2003. *Learning to Write: Error Analysis Applied*. Universidad de Alcalá: Servicio de Publicaciones.

*Corresponding author*

Zeltia Blanco-Suárez

Faculty of Education (Department of Philology)

University of Cantabria

Avenida de los Castros s/n

39005 Santander

e-mail: zeltia.blanco@unican.es

received: February 2020

accepted: March 2020

# The *Toledo Teacher Trainees* corpus (TTT): Bridging the gap between students' narratives and corpus linguistics

Fátima Faya-Cerqueiro<sup>a</sup> – Gema Alcaraz-Mármol<sup>b</sup>  
University of Santiago de Compostela<sup>a</sup> / Spain  
University of Castilla La Mancha<sup>b</sup> / Spain

**Abstract** – In recent decades a few research methods have resorted to L2 learners in order to analyse several aspects aiming at methodological improvements. One of them is corpus linguistics, which has largely contributed to the study of language production from a quantitative perspective. A very different one has been the compilation of perceptions of the L2 learning process using ‘narrative inquiry’ and qualitative methods of analysis. However, scholars have not addressed the combination of both methods. In this proposal we examine their main individual features and offer an interwoven line of research, applying the quantitative approach of corpus linguistics to the genre of language learning narratives. Thus, we present a new corpus of L2 learners’ perceptions and provide detailed information on its structure, compilation and categorisation. The interdisciplinary status of this proposal will enable the exploration of new research possibilities that can ultimately benefit the teaching-learning process.

**Keywords** – narrative inquiry; L2 learning narratives; corpus linguistics; teacher trainees; corpus-based perceptions

## 1. INTRODUCTION<sup>1</sup>

The twenty-first century has witnessed a growing interest in how second language (L2) students conceive their own learning process. The interest and effort to improve this process on the part of the teaching community and linguists is not new. Nonetheless, this interest has traditionally left aside the reflections and perceptions of the learners themselves, focusing on more objective aspects such as the product of the learning process. It is precisely in the light of these facts that learner corpora were born at the beginning of the 1990s, centred on the analysis of the written and oral productions of L2

---

<sup>1</sup> The authors are grateful to the research group CACLE (*Comunicación, Aprendizaje y Competencias en Lengua Extranjera*) for their generous support.

students. However, learner corpora generally focus on the learners' L2 output, not on the learning process itself. In this sense, Pavlenko (2007: 163) affirmed:

In the past decade, language memoirs, linguistic autobiographies, and learners' journals and diaries have become a popular means of data collection in applied linguistics. It is not always clear however how one should go about analysing these data.

These memoirs, autobiographies and journals allow a new approximation to the learning process from the learners' standpoint itself. These elements offer valuable data for researchers, teachers and teacher trainers, which lead to the potential improvement of teaching methodologies.

However, despite offering some directions for systematic analysis, the focus of these studies is eminently qualitative. Pavlenko's statement reveals there is still room to pose different research questions and adopt new perspectives when tackling this genre. In this sense, corpus linguistics may constitute a methodological framework where data, such as the content of these narratives, can be approached from a perspective more in line with a quantitative standpoint. What is suggested here is the use of corpus within a broader scope and to go beyond L2 output analysis, paying attention to aspects which have more to do with the exercise of reflection and self-report about the learning process itself.

In what follows, we present the *Toledo Teacher Trainees* corpus (TTT), based on L2 learners' narratives. We aim at building a bridge between two approaches, namely corpus linguistics and learners' perceptions, in an attempt to address the data taken from the latter through the method promoted by the former. We will first describe the theoretical framework behind our proposal (cf. Section 2). This framework explores the contribution of L2 learning narratives to the perception of the teaching learning process, as well as the potential role of corpus linguistics in Second Language Acquisition in general and in the analysis of these narratives in particular. We pay particular attention to learner corpus research, a relatively recent area of study that shares an interest in learners' language learning process. In fact, a comparison between learner corpora, narratives and our own methodology is established. Section 3 provides a detailed description of our corpus, including compilation methods, participants' profile, main steps taken for analysis and categorisation. Our paper finishes with perspectives and research directions regarding potential benefits and advantages that our corpus offers for future studies (cf. Section 4).

## 2. PREVIOUS RESEARCH

Over the last 35 years, corpus linguistics research has dealt with the description of different types of registers, among which we can find informal conversation, journals, reports, female and male language or even dialects. As regards language learning and teaching, more and more researchers highlight how corpus linguistics can help this discipline (Aijmer 2002; Reppen 2010; Frankenberg-Garcia *et al.* 2011). In Römer's (2011: 205) words, "in the field of applied linguistics, more and more researchers and practitioners treasure what corpus linguistics has to offer to language pedagogy." This particular field within corpus linguistics, dealing with the collection of linguistic evidence from non-native speakers, is known as learner corpus studies.

The origins of learner corpora are found in the late 1980s, but they did not really take root as a discipline until the beginning of the 1990s. A learner corpus brings together corpus linguistics and language learning, aiming at providing descriptions of learner language, and offering new perspectives on second language learning, and sometimes changing the ones already established. It helps us gain better understanding of how languages other than the first language (L1) are learned (Ortega 2009). Paquot and Plonsky (2017: 2) define it as an "interdisciplinary enterprise which sits its crossroads between a variety of disciplines."

The focus of a learner corpus is to be found in two types of studies. The first type pursues the analysis of L2 output in the written or spoken form, exploring several linguistic patterns of different nature within the learners' interlanguage. The linguistic patterns of interest for research include syntactic complexity (Vyatkina 2013), frequency and correctness of vocabulary (Laufer and Waldman 2011) and even pragmatics (Chen 2010; Polat 2011). The second type delves around what is called 'Error Analysis', which helps us understand the development of the learning process (Granger 2002). Among this second category, we find Divsar and Heydari (2017) and Botley *et al.* (2007). The former analyse several categories of errors in Iranian English as a Foreign Language learners' essays, whereas the latter pay attention to spelling.

The range of learner corpora that exist nowadays varies in number and category. The *Cambridge Learner Corpus* (Nicholls 2003) and the *International Corpus of Learner English* (Granger *et al.* 2009) are among the most ambitious, with data banks of learners from different L1 backgrounds and millions of words. These two corpora use the written language as a medium. However, others such as the *College Learner Spoken English*

*Corpus* (Yang and Wei 2005), the *Corpus of the National Institute of Information and Communications Technology for Japanese Learners of English* (Izumi *et al.* 2004), or the *Corpus of Young Learner Interlanguage* (Housen 2002) rely on spoken language. They are usually smaller than the written ones but equally relevant.

Learner corpus linguistics has adopted this standpoint based on the analysis of L2 learners' output, paying attention to the learners' interlanguage and errors in order to contribute to the improvement of language teaching methodology. We share this same aim but, in our case, the object of analysis is not the L2 production, but the learners' perceptions on their own L2 learning process.

These reflections are the object of research of what it is known as 'narrative inquiry.' Narrative inquiry, broadly defined by Barkhuizen *et al.* (2014: 3) as "an established umbrella term for research involving stories," became a prominent research method in social science at the turn of the twenty-first century (Clandinin and Connelly 2000). The recognition of narratives as valid data and the need of qualitative approaches to cover personal experiences were major assets for the development of this methodology (Huber *et al.* 2013: 217). Some authors have claimed the relevance of narratives together with a need for systematicity in their study (Atkinson and Delamont 2006), and several proposals regarding design, collection and analysis have actually provided a consistent foundation for research (Pavlenko 2007; Riessman 2008; Wells 2011). Nowadays, this "interdisciplinary method that views lives holistically" (Marshall and Rossman 2014: 157) is a well-established and productive approach encompassing work from different areas, as attested by the variety of disciplines included in first volumes devoted to narrative inquiry (Clandinin 2007) or by the journal that bears the same name.

Within social science, the field of education, probably due to its broad scope, has adopted this methodology for an array of studies, where narrative inquiry can provide a better understanding of teaching and teacher education, classroom practice, classroom management or curriculum design, among others (Johnson and Golombek 2002; Craig 2011; Kitchen *et al.* 2011; Huber *et al.* 2013). As regards L2 learning the most common areas of research using learners' narratives include motivational issues (cf. Thompson and Vásquez 2015; Farahani *et al.* 2019), identity (Early and Norton 2012; Benson *et al.* 2013) and also more specific perceptions on given topics, like the native speaker myth, related to the lack of confidence perceived by non-native language teachers (Reis 2011), test-takers' perspectives (Rajendram *et al.* 2019; Sinclair *et al.* 2019) or assessment

(Franco 2020). Although most narratives deal with general issues regarding L2 previous experiences, research can also include more specific or tailored narratives, such as study-abroad experiences (Benson *et al.* 2013) or learners' perceptions of teachers (Oxford 2001). The analysis of teachers' or learners' narratives, in particular, has also demonstrated its relevance for practitioners, future students and learners to reflect upon their experiences. Similarly, research on educational stakeholders' perspectives can be used to implement or improve educational practices, being thus a cyclic process when teaching and learning perceptions can eventually change teaching and learning attitudes.

In the L2 context, narrative inquiries have mostly been approached through face-to-face interviewing processes, which makes data gathering a costly and time-consuming method of compilation requiring interaction between the learner and the researcher. Moreover, the number of participants in L2 narrative inquiry research is generally limited due to this procedure, since when using interviews the time devoted to compilation and transcription can equal a low number of participants in those studies. Quantitative approaches to L2 self-perceptions are virtually non-existent. Yet, a study by Baker and MacIntyre (2000), which examines a French immersion programme in Canada, could be in the line of a quantitative perspective. As a way to complement their survey, based largely on quantitative research methods and supported by statistical tests, they asked students to describe a positive or negative experience about speaking French. Although they refer to this part of the study as qualitative, it does include percentages of different lexical items regarding the experiences in the immersion and non-immersion groups.

Table 1, below, shows the differences and similarities found in traditional learner corpora, corpus of learners' perceptions and narrative inquiry, as applied to L2 learning experiences. The comparison is established upon eight features. The three fields of study share some features such as the participants' profile, the format and the possible application of results. By contrast, important differences are observed, namely the scope, the number of participants and the types of results obtained. Traditional learner corpora are mainly concerned with L2 production, while the other two analyse self-reflection on the language learning process, which can be expressed in the learners' L1 or in L2, depending mainly on language proficiency.

	<b>Traditional learner corpus</b>	<b>Corpus of learners' perceptions</b>	<b>Narrative inquiry</b>
<b>Participants' profile</b>	L2 learners	L2 learners	L2 learners
<b>Number of participants</b>	Representative	Representative	Case studies
<b>Format</b>	Oral/written	Written	Oral/written/multimodal
<b>What do they study?</b>	Linguistic data	Metalinguistic data	Metalinguistic data
<b>Language</b>	L2	L1	L1/L2
<b>Aim</b>	Interlanguage / errors	Learning experience	Learning experience
<b>Application</b>	Methodology	Methodology	Methodology
<b>Type of results</b>	Quantitative	Quantitative	Qualitative

Table 1: Differences and similarities between traditional learner corpus, corpus of learners' perceptions and narrative inquiry

The review of the literature reveals a gap, as the combination of both elements – corpus linguistics and narrative inquiry – has not yet been tackled. Thus, our proposal differs from the abovementioned works in two main aspects. First, research on learners' perceptions has been developed with a small number of participants, and generally focused on a particular element within the learning process. Second, those studies adopt a qualitative perspective, maybe partly due to the small number of participants. We consider, therefore, that corpus linguistics can open new lines of research as regards narrative inquiry and students' perceptions, adopting a more quantitative perspective and with a significantly higher number of participants.

### 3. *TOLEDO TEACHER TRAINEES* (TTT): A CORPUS OF L2 PERCEPTIONS

#### 3.1. *Compilation, structure and main principles*

The *Toledo Teacher Trainees* corpus (TTT) aims to be at the crossroads between learner corpora and L2 narratives, as it provides a collection of learners' perceptions which has been designed to make it a representative source of information for different research purposes. The participants were 354 future teachers of Primary and Infant Education from the Faculty of Education in Toledo, Spain. The four academic years in these degrees are represented in our sample with 162 students in their first year, 123 in their second, 42 in their third and 27 in their fourth. After completion of their second year, students achieve a B1 level, while the last two years belong to a special language training group that grants a B2 level. Most of the participants are taking a Degree in Primary Education (269) and the rest study a Degree in Infant Education (85). Their ages range from 18 to 45 with a

mean of 20.0 years. Regarding gender they fall into the following categories: 245 female, 109 male. All of them declare Spanish as their native language except for four, three of whom spoke Romanian and one Valencian as L1. Only 14% of them have not studied a third language (50). Among the 354 students, 147 have taken part in a bilingual CLIL programme (English/Spanish) at some stage before university.

The compilation of the corpus was done through an *ad-hoc* online questionnaire. Learners had to fill in the questionnaire individually. Different sessions were organised for each group to enable simultaneous participation without time limit and they were supervised by their language teacher and/or researcher to answer possible questions and to provide information on the purposes of the compilation. The questionnaire was not shown to the participants previously although they were briefly informed about its main features. All the contributions were compiled during the same academic year.

The questionnaire consisted of two main parts. The first part contained questions regarding the profile of the participants such as age, gender, degree, L1, additional languages and educational background. As for the second part it included specific items concerning their experience in the different pre-university stages. Participants had to answer open questions with no word limit and a closed question where they were required to write three keywords related to their experience, which could summarise their perceptions of each stage. Open questions comprehended different aspects of their academic life such as teachers, materials and resources, methodology, contents and assessment process.

The corpus comprises around 170,000 words and includes participants' answers to the open questions regarding each of the mentioned educational stages as well as the keywords associated to the different school periods.

All the questions were written in Spanish and this language was also required for their responses. The reason why their L1 was adopted was to guarantee they could express themselves without linguistic limitations. This is common practice in studies that focus on metacognitive or metalinguistic features and also in L2 narratives (Pavlenko 2007: 172):

in studies of subject and life reality where the speakers' L2 proficiency is low and the L1 is shared with the researcher, the choice of L1 as the language of data collection is justified.

Accordingly, Farahani *et al.*'s (2019) research B1 level participants use their L1, Persian, when writing language learning histories.

### 3.2. *Our approach*

The present study is part of a wider research project that intends to explore perceptions about the L2 teaching-learning process of university students in earlier stages. The corpus analysis is developed in two stages. First, we focus on the keyword list of each stage, a phase which includes systematisation and analysis of these words, which are classified into broad semantic categories. This first step enables the comparison of different patterns identified in the keywords. A second stage will resort to this classification for the analysis of the answers to the open questions.

*AntConc* (Anthony 2019) was used to alphabetically organise and count all the keyword instances, which follow the clearance of non-valid answers, such as whole clauses, prepositions and other non-content words since our main aim was to see students' major topics of interests. The selection focused on nouns, verbs and adjectives. For practical reasons all the tokens were unified into types, so these types included masculine and feminine adjectives (*cómodo/a*, 'comfortable') and word categories with the same lemma (*diferencia/diferente* 'difference/different'; *aprendizaje* 'learning'/verb forms of *aprender* 'to learn').

Each type was classified according to broad semantic fields in different stages. For this purpose there was a progression from narrower sets or number of categories to broader ones. Three evaluators were simultaneously immersed in this process, which consisted of assigning a label to each token or more than one in an initial stage. In the first explorations categories amounted to near 30 and were finally reduced to 11.

Umbrella terms were applied to categories. The idea was to present them as neutral concepts as possible. In this way, they would become comprehensible enough to admit all the different possibilities within a continuum. Some of the final categories are 'emotion/feelings', 'complexity', 'innovation' or 'usefulness.' They show a great variability regarding number of members, token/type ratio, range of word categories and also the type of concepts that they define (abstract vs. concrete ideas).

However, semantic categories are not always absolute terms and blurred lines between categories do occur. Simon-Vandenberg and Aijmer (2008: 12) know the

difficulty when defining “the semantic field a priori and there is no obvious and well-defined number of dimensions structuring the field.” This idea contrasts with the mosaic conception of lexical fields, where components fit nicely. In fact, Geeraerts *et al.* (1994) warn that there might be overlapping between different fields given their fuzzy boundaries. Accordingly, non-standard characteristics are established, which are far from the traditional established ones. In those cases, agreement on the category which better fits the word is supported by dictionaries and thesaurus.

Far from theorising on semantics itself, semantic concepts are employed to create useful categories for the purposes of our research, always within coherence. The literature offers a plethora of definitions that remark the relationship of words or lemmas that share semantic features. According to Mackey (1965: 76), a semantic field is:

made up of basic key-words, which command an army of others. The semantic area may be regarded as a network of hundreds of associations, each word of which is capable of being the centre of a web of associations radiating in all directions.

Words with common semantic associations can be categorised as members of such a network. As noted by Kittay and Lehrer (1992: 3), “words applicable to a common conceptual domain are organised within a semantic field by relations of affinity and contrast (e.g., synonymy, hyponymy, incompatibility, antonymy, etc.).” Following this idea, the notion of semantic field is understood in a broad sense here, comprising different hierarchical associations. Core members of each semantic field or category are easily identified, while more peripheral members require the support of dictionary definitions. In some cases, definitions of a given entry are insufficient and further searches of words within the definition are needed.

The category *cambio* ‘change’, which is among the shortest categories, can serve as an example. It would include not only ‘change’ alone, but also ‘lack of change’ and all the different options within the cline ‘change-lack of change.’ Thus, the following members have been identified: *ágil* (‘agile’), *cambio* (‘change’), *diferencia* (‘difference’), *dinámico* (‘dynamic’), *diverso* (‘diverse’), *evolución* (‘evolution’), *monótono* (‘monotonous’) and *repetitivo* (‘repetitive’).

#### 4. PERSPECTIVES AND RESEARCH DIRECTIONS

As already noted, narrative inquiry can bring about immediate benefits for education students. Reflecting on previous experiences as learners may also lead future practitioners to become aware of the kind of teacher they want to be, the methodological approaches to use and other aspects concerning their teaching practice. At the same time, this narration of their L2 learning experiences can provide researchers with valuable data to detect weaknesses and strengths regarding L2 teaching and aspects that might influence L2 learning in various ways. These data can turn into suggestions for action at different levels.

In our opinion, corpus linguistics can enrich the potential of these narratives, providing a new approach to their analysis. The compilation of a structured corpus with different variables and the introduction of quantitative methods to the genre of L2 narratives will enable sharper research questions and more solid generalisations on how the teaching-learning process of different school periods is perceived. The approach to categorisation presented above is just an example of the possibilities offered by the corpus. Some of these possibilities for further research have already been suggested. Similarly, the specific and precise categorisation of lexical items will facilitate corpus searches in the narratives, not only between elements of the corpus itself, but beyond. Although in the present article only one category has been sketched, there are more semantic categories which have been established on the basis of the keyword analysis. In turn, these categories can be completed with new members identified in the corpus beyond the keywords. These semantic categories can also serve as a starting point for carrying out searches for specific terms of interest related to the teaching and learning process.

The wide range of criteria taken into account for data collection will allow us to make systematic comparisons including variables such as gender, linguistic background, educational stage and even aspects on which dense information has been obtained such as teachers, assessment or resources, which stand as object of reflection for our participants. Moreover, comparisons can be made on the basis of the different academic years to which the participants belong. Adopting an apparent-time perspective and having students from all years in the same degree, we have the opportunity to observe the evolution of perceptions from the first to the last year of their university stage.

In this first contribution to corpus-based perceptions we have introduced our compilation approach. Nevertheless, our corpus offers a wide range of possibilities yet to be explored, which can be of interest to linguists, teachers and teacher trainers. However, a whole corpus remains to be exploited beyond the keywords, which are the easiest part to identify within the corpus, but the most constrained as well.

The specific nature of our corpus encourages us to contrast our results with other specific or more general corpora. Some of these comparisons should be taken with caution, since our corpus is framed within a highly specific context, where most students speak one and the same L1. Besides, most of them belong to the same geographical area in the province of Toledo. Therefore, some evidence from our corpus cannot be extrapolated. However, in spite of the possible limitations derived from context, the corpus structure, its longitudinal and transversal nature will allow to establish comparisons with similar corpora in different contexts of a multilingual nature.

Not only do the possibilities for comparison remain intracorporal, but they can also be found intercorpus. Indeed, one could study aspects of interest to the field of learner corpus but from a perception perspective. That is, as mentioned above, errors are one of the most explored aspects in learner corpora, and these are analysed from a strict perspective of linguistic correction. In our corpus, errors may constitute an element for reflection among the participants, being thus treated from a metalinguistic point of view. Similarly, we could deal with aspects related to oral and written production, within a framework of reflection on the learning process itself.

On the other hand, comparisons could also be carried out with aspects tackled in narrative inquiry research, such as aspects of identity, self-concept and attitude towards the L2. The corpus perspective would provide a quantitative basis in this sense. Furthermore, the potential of this corpus goes beyond the quantitative, since it offers material that can be qualitatively analysed through subsequent interviews with a small group of participants, taking objective data as a starting point.

Therefore, and in spite of the specificity that characterises the corpus presented here, we consider that the study points to the interdisciplinary value of second language research, particularly in the field of perceptions, where the systematicity offered by corpus methodology paves new paths of research.

## 5. REFERENCES

- Anthony, Laurence. 2019. *AntConc*. Tokyo: Waseda University.
- Aijmer, Karin. 2002. Modality in advanced Swedish learners' written interlanguage. In Sylviane Granger, Joseph Hung and Stephanie Petch-Tyson eds. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins, 55–76.
- Atkinson, Paul and Sara Delamont. 2006. Rescuing narrative from qualitative research. *Narrative Inquiry* 16/1: 164–172.
- Baker, Susan C. and Peter D. MacIntyre. 2000. The role of gender and immersion in communication and second language orientations. *Language Learning* 50/2: 311–341.
- Barkhuizen, Gary, Phil Benson and Alice Chik. 2014. *Narrative Inquiry in Language Teaching and Learning Research*. New York: Routledge.
- Benson, Phil, Gary Barkhuizen, Peter Bodycott and Jill Brown. 2013. *Second Language Identity in Narratives of Study Abroad*. Basingstoke: Springer.
- Botley, Simon, Faizal Hakim and Doreen Dillah. 2007. Investigating spelling errors in a Malaysian learner corpus. *Malaysian Journal of ELT Research* 3/1: 74–93.
- Chen, Hsin-I. 2010. Contrastive learner corpus analysis of epistemic modality and interlanguage pragmatic competence in L2 writing. *Arizona Working Papers in SLA and Teaching* 17/1: 27–51.
- Clandinin, D. Jean ed. 2007. *Handbook of Narrative Inquiry: Mapping a Methodology*. Los Angeles: Sage Publications.
- Clandinin, D. Jean and F. Michael Connelly. 2000. *Narrative Inquiry: Experience and Story in Qualitative Research*. San Francisco: Jossey-Bass.
- Craig, Cheryl. 2011. Narrative inquiry in teaching and teacher education. *Narrative Inquiries into Curriculum Making in Teacher Education* 13: 19–42.
- Divsar, Hoda and Robab Heydari. 2017. A corpus-based study of EFL learners' errors in IELTS essay writing. *International Journal of Applied Linguistics and English Literature* 6/3: 143–149.
- Early, Margaret and Bonny Norton. 2012. Language learner stories and imagined identities. *Narrative Inquiry* 22/1: 194–201.
- Farahani, Ali Akbar Khomeijani, Abbas Ali Rezaee and Robabeh Moshtaghi Zonouz. 2019. Motivational trajectories in language learning: Evidence from highly-motivated English as a foreign language learners. *Electronic Journal of Foreign Language Teaching* 16/2: 281–299.
- Franco, Ashleigh. 2020. Not all Finns think alike: Varying views of assessment in Finland. *International Education Studies*. 13/1: 1–10.
- Frankenberg-Garcia, Ana, Lynne Flowerdew and Guy Aston. 2011. *New Trends in Corpora and Language Learning*. London: Continuum.
- Geeraerts, Dirk, Stefan Grondelaers and Peter Bakema. 1994. *The Structure of Lexical Variation: Meaning, Naming and Context*. Berlin: Mouton de Gruyter.
- Granger, Sylviane. 2002. A bird's-eye view of learner corpus research. In Sylviane Granger, Joseph Hung and Stephanie Petch-Tyson eds. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins, 3–33.
- Granger, Sylviane, Estelle Dagneaux, Fanny Meunier and Magali Paquot. 2009. *International Corpus of Learner English*. Louvain-la-Neuve: Presses Universitaires de Louvain.

- Housen, Alex. 2002. A corpus-based study of the L2 acquisition of the English verb system. In Sylviane Granger, Joseph Hung and Stephanie Petch-Tyson eds. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins, 77–116.
- Huber, Janice, Vera Caine, Marilyn Huber and Pam Steeves. 2013. Narrative inquiry as pedagogy in education: The extraordinary potential of living, telling, retelling and reliving stories of experience. *Review of Research in Education* 37/1: 212–242.
- Izumi, Emi, Kiyotaka Uchimoto and Hitoshi Isahara. 2004. *Nihonjin 1,200-nin no Eigo Speaking Corpus (A speaking corpus of 1,200 Japanese learners of English)*. Tokyo: ALC Press.
- Johnson, Karen E. and Paula R. Golombek. 2002. *Teachers' Narrative Inquiry as Professional Development*. Cambridge: Cambridge University Press.
- Kitchen, Julian, Darlene Ciuffetelli Parker and Debbie Pushor eds. 2011. *Narrative Inquiries into Curriculum Making in Teacher Education*. Bingley: Emerald.
- Kittay, Eva Feder and Adrienne Lehrer. 2012. Introduction. In Adrienne Lehrer, Eva Feder Kittay and Richard Lehrer eds. *Frames, Fields and Contrasts: New Essays in Semantic and Lexical Organization*. New York and London: Routledge.
- Laufer, Batia and Tina Waldman. 2011. Verb-noun collocations in Second Language writing: A corpus analysis of learners' English. *Language Learning* 6/2: 647–672.
- Mackey, William F. 1965. *Language Teaching Analysis*. London: Longman and Indiana University Press.
- Marshall, Catherine and Gretchen B. Rossman. 2014. *Designing Qualitative Research*. Los Angeles: Sage Publications.
- Nicholls, Diane. 2003. The Cambridge Learner Corpus– error coding and analysis for lexicography and ELT. In Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery eds. *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster University: University Centre for Computer Corpus Research on Language, 572–581.
- Ortega, Lourdes. 2009. *Understanding Second Language Acquisition*. London: Hodder Education.
- Oxford, Rebecca L. 2001. The bleached bones of a story: Learners' constructions of language teachers. In Michael P. Breen ed. *Learner Contributions to Language Learning: New Direction in Research*. London: Longman, 86–111.
- Paquot, Magali and Luke Plonsky. 2017. Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research* 3: 61–94.
- Pavlenko, Aneta. 2007. Autobiographic narratives as data in applied linguistics. *Applied Linguistics* 28/2: 163–188.
- Polat, Brittany. 2011. Investigating acquisition of discourse markers through a developmental learner corpus. *Journal of Pragmatics* 43/15: 3745–3756.
- Rajendram, Shakina, Jeanne Sinclair and Elizabeth Larson. 2019. International graduate students' perspectives on high-stakes English tests and the language demands of Higher Education. *Language and Literacy* 21/4: 68–92.
- Reis, Davi Schirmer. 2011. "I'm not alone:" Empowering non-native English-speaking teachers to challenge the native speaker myth. In Karen E. Johnson and Paula R. Golombek eds. *Research on Second Language Teacher Education*. New York: Routledge, 45–63.
- Reppen, Randi. 2010. *Using Corpora in the Language Classroom*. Cambridge: Cambridge University Press.

- Riessman, Catherine Kohler. 2008. *Narrative Methods for the Human Sciences*. Los Angeles: Sage Publications.
- Römer, Ute. 2011. Corpus research applications in Second Language teaching. *Annual Review of Applied Linguistics* 31/1: 205–225.
- Simon-Vandenberg, Anne-Marie and Karin Aijmer. 2008. *The Semantic Field of Modal Certainty: A Corpus-Based Study of English Adverbs*. Berlin: Walter de Gruyter.
- Sinclair, Jeanne, Elizabeth Jean Larson and Shakina Rajendram. 2019. “Be a machine:” International graduate students’ narratives around high-stakes English tests. *Language Assessment Quarterly* 16/2: 236–252.
- Thompson, Amy S. and Camilla Vásquez. 2015. Exploring motivational profiles through language learning narratives. *The Modern Language Journal* 99/1: 158–174.
- Vyatkina, Nina. 2013. Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus. *The Modern Language Journal* 97/1: 11–30.
- Wells, Kathleen. 2011. *Narrative Inquiry*. Oxford: Oxford University Press.
- Yang, Huizhong and Naixing Wei. 2005. *The Construction of and Research on the COLSEC*. Shanghai: Shanghai Foreign Language Education Press.

*Corresponding author*

Fatima Faya-Cerqueiro  
 University of Santiago de Compostela  
 Department of Applied Didactics  
 E-15782 Santiago de Compostela  
 Spain  
 e-mail: fatima.faya@usc.es

received: January 2020

accepted: May 2020

Review of Fanego, Teresa and Paula Rodríguez-Puente eds. 2019. *Corpus-based Research on Variation in English Legal Discourse*. Amsterdam: John Benjamins. ISBN: 978-9-027-20235-2. <https://doi.org/10.1075/scl.91>

Christopher Williams  
University of Foggia / Italy

### 1. SUMMARY

This volume is part of the John Benjamin *Studies in Corpus Linguistics* series, the second specifically analysing legal language, after the one edited by Laura Mori (2018).

In *Corpus-based Research on Variation in English Legal Discourse* (hereafter abbreviated as *CRVELD*), except for Chapter 1 which outlines the main themes of modern legal discourse research and summarises the other chapters, the volume is divided into two parts, each comprising five chapters. Part One focuses on ‘Cross-genre and cross-linguistic variation’, while Part Two is concerned with ‘Diachronic variation’.

As Teresa Fanego and Paula Rodríguez-Puente, respectively from the University of Santiago de Compostela and the University of Oviedo, point out in the Acknowledgements, the idea for this volume arose when the two editors were involved in compiling the *Corpus of Historical English Law Reports 1535–1999*.

The opening chapter presents an overview of legal discourse studies, starting with Bhatia’s (1987) observation about the dramatic rise in the number of studies in legal language. Since then legal discourse studies have grown apace, and the editors offer a fascinating synopsis of the various strands that have evolved, highlighting three relatively innovative areas, i.e. *FASP* (*fiction à substrat spécialisé*) which, in the legal sphere, focuses on law-related fiction, including TV series and films; the impact of plain language on legal discourse; and forensic linguistics, another multi-faceted field in rapid expansion.



The authors then turn to the question of register and genre perspectives on legal discourse, underlining the importance of works by Biber (e.g., Biber 1988) and Swales (e.g., Swales 1990) when applied to the legal sphere. It is in this broader context, together with the exponential growth of corpus linguistics, that the explosion of interest in analysing legal English and other specialised domains has to be understood.

Taxonomies of legal text types are briefly explored, again starting with Bhatia (1987, 1993) and his subdivisions of the various types of spoken texts and written texts, later supplemented by Tiersma (1999) and Šarčević (2000). The authors then briefly survey legal discourse and (historical) pragmatics. Although the spoken language of the past cannot be investigated through direct observation, they affirm that methodologies have been refined so that it is now “possible to get an approximate picture of the spoken language of past centuries” (9). The rest of the chapter is devoted to summaries of the contributions making up the rest of the volume, based on Biel’s (2010: 4–5) classification of four major ‘trajectories’: external variation, internal variation, temporal variation, and cross-linguistic variation.

In her chapter “English and Italian land contracts: A cross-linguistic analysis,” Giuliana Diani observes how Italian contracts differ from their English counterparts: the recitals section is generally absent in Italian texts; the organisation and layout of the section on operative provisions is different; and there is more punctuation in Italian texts.

Comparing linguistic features, the author highlights the frequency in both corpora of lexical repetition and the paucity of personal pronouns. She remarks that only masculine pronouns are employed in English texts, and that both corpora repeatedly display compound adverbs and anaphoric expressions. She also notes the “abundance of complex prepositions and other idiomatic and semi-idiomatic sequences with the structure preposition + noun + preposition” (33) in both languages.

The author compares binomial and multinomial expressions, observing that English contracts show a greater range and frequency. A further difference concerns the omission in English of the article before nouns denoting the party’s functional role (e.g. *Purchaser*, *Vendor*), unlike Italian. She asserts that syntactic complexity and excessive

wordiness are common to both corpora. Passive constructions, Diani notes, are considerably more frequent in the English corpus.

Regarding tense and modality, there is an “enormous disparity in the frequency of deontic modals” (39). Diani hypothesises that Italian contracts adopt a variety of “other deontically-charged devices” (43) such as *obbligare/arsi* or *promettere*. She also observes that deontically-charged nominals such as *diritto* or *facoltà* are frequently used. However, she affirms that possibly the most important deontic device in Italian contracts is the frequency of the simple present indicative and the future form of verbs.

She concludes that although both sets of texts have a shared communicative function, the language of Italian contracts is generally less formulaic and exhibits greater variation.

In her chapter Cristina Lastres-López examines conditionals in spoken courtroom discourse in English, Spanish and French, and conditionals in spoken parliamentary discourse in English and French. In differentiating between courtroom discourse and parliamentary discourse, the author affirms that while “the former clearly qualifies as a ‘purely’ legal register, the latter represents a more hybrid text type” (57).

Analysing the results relating to courtroom discourse, Lastres-López observes that conditional clauses in English can be introduced by various markers besides *if*, whereas in Romance languages, using “conditionals with markers other than *si* is very marginal” (61). As for metafunctions, ideational conditionals are easily the most common in all three corpora. Interpersonal conditionals are less common, and largely restricted to the English corpus.

In terms of the semantic type of condition, real conditions prevail in all three corpora, followed by potential conditions and unreal conditions. As for modality, in English the most frequent modals are *would*, *may* and *will*, whereas in French and Spanish the meanings expressed by modals “are encoded by verbal endings on the lexical verb” (64).

In investigating conditionals in parliamentary discourse, Lastres-López restricts her inquiry to 500 *if*-clauses and 500 *si*-clauses, observing that both languages exhibit a clear preference for ideational conditionals.

Regarding the semantic type of condition, the results confirm the same sequence as in courtroom discourse, with real conditions dominating in both languages. In terms

of modality, over 67 per cent of conditionals in English occur with modal verbs, whereas in French only 21 per cent of apodoses contain a modal.

Ruth Breeze's chapter "Part-of-speech patterns in legal genres: Text-internal dynamics from a corpus-based perspective" examines four corpora –academic texts, case law, legal documents and legislation– all related to business law. After outlining some methodological issues, the author discusses the results, highlighting "some similarities across all the corpora, particularly concerning the frequency of possessive nouns" (83), and pointing to "a greater tendency to use the Saxon genitive in legal writing overall than in the construct of general English" (84). However, she also finds "some major differences between the four corpora, suggesting that they fall into two groups: Academic and Cases, on the one hand, and Documents and Legislation, on the other" (84).

Breeze then focuses on the key features across the corpora, observing the higher frequency of plural nouns compared with the BNC reference corpus. Regarding singular possessive nouns, the author highlights the striking result of the Legislation corpus where most instances consist of the same word, i.e. *company's*.

The author observes that third person singular present tenses are key in the Academic and Cases corpora. She also notices the salience of "instances where *that* is part of a linking expression with a connecting function" (87).

As for the key features in Documents and Legislation, Breeze remarks on the high keyness of coordinating conjunctions. Another key feature is the frequency of past participles. The author affirms that "[t]he preference for using the passive is the main explanation for the frequency of participles" (94).

Breeze notes the frequency of particular modal verbs in Documents and Legislation, above all *shall*. She affirms that the high frequency of *must* in Legislation provides evidence that drafters "have adopted some of the principles of plain English" (96), whereas legal documents are generally more conservative, "hence their greater tendency to preserve *shall*" (96).

The chapter by Randi Reppen and Meishan Chen "explores variation in spoken courtroom language across time, registers, and varieties using three-word lexical bundles (Biber *et al.* 1999)" (105). Based on Culpeper and Kytö's (2010: 103–141) account of lexical bundles in Early Modern English (EModE) trials (1560–1760) and

Present Day English (PDE) trials (1993), the authors compare the bundles with those in the 1994 O.J. Simpson trial.

Reppen and Chen first describe the corpora and the methodology used: in the O.J. trial this includes a description of the sub-registers. In comparing three-word bundles, they list “the fifty most frequent bundles found in the O.J. Simpson trial along with the bundles that also occurred in the EModE and PDE corpora” (109). 76 per cent of the bundles are specific only to the O.J. trial, often referring to particular places, people, or times. The authors point to the high frequency of bundles in O.J. containing *you*, whereas “the EModE and PDE bundles had a greater use of ‘speaker-centered’ bundles” (111).

The most frequent bundles common to all three corpora are *at that time*, *what did you*, and *one of the*, which “share the discourse goal of providing information and specifics” (113).

Analysing sub-registers in the O.J. corpus, the authors find that in opening statements bundles are frequently evidential and sensory-related, which “has the effect of guiding the jurors’ thoughts” (115). Interrogatives feature repeatedly in direct examination bundles, while cross-examination bundles tend to highlight the challenge to the credibility of witnesses. Surprisingly, the authors provide little analysis of bundles data relating to the fourth sub-register listed, i.e. closing argument.

Stanisław Goźdź-Roszkowski’s chapter explores stance construction in legal writing by analysing how the Noun *that*-pattern is used in academic journals and judicial opinions. He observes that most studies on stance in the legal domain to date have focused on judicial discourse, unsurprisingly given the importance of stance or evaluation for judicial argumentation.

The analysis draws on three different datasets: the *Academic Journals Corpus* (AJC), the *Corpus of Judicial Opinions* (CJO) and the *British Law Report Corpus* (BLaRC). Noun Complement structures are investigated where head nouns “take a nominal complement in the form of *that*-clause” (129). The nouns were then classified into three types governing *that*-clauses: epistemic nouns, attitude or perspective nouns and communication (non-factual) nouns.

The author notes that most nouns fall into the category of epistemic nouns, especially those typically associated with certainty or marking likelihood, while those

signalling attitude or perspective or corresponding to the category of communicative nouns are less common. He observes that “there are considerable differences between the three corpora, and particularly between the AJC and the two judicial corpora with regard to epistemic nouns indicating certainty and communication nouns” (130).

Goźdz-Roszkowski then focuses on two nouns –*fact* and *conclusion*– that he considers “particularly revealing in terms of radically different distributions across the three corpora and the various ways in which they are used by legal writers” (133). He observes that *fact that* is especially salient in judicial writing.

The author concludes that while the Noun *that*-pattern is generally used in a similar way in both academic and judicial writing, “academic writing differs from judicial writing in that it sometimes uses nouns with a *that*-clause complement to construct a stance in more neutral ways intended to appear to be objective and impersonal” (143).

Part II on ‘Diachronic variation’ opens with Douglas Biber and Bethany Gray’s paper which considers how far law reports have adopted linguistic innovations discernible in other written registers, concluding that law reports belong to the so-called *uptight* written registers (Hundt and Mair 1999), as distinct from those *agile* registers that have drifted towards a more colloquial style.

The authors observe that, unlike academic prose, law reports have remained relatively constant because they have always been written primarily for specialist readers, and “have maintained their primary communicative purposes of documenting the facts of a legal case” (152) and explaining the grounds for the judgment.

The authors investigate law reports from 1700 to 1999 using the *Corpus of Historical English Law Reports* (CHELAR; see Fanego *et al.* 2017), comparing them with the registers of fiction, newspaper articles, and science research articles. They observe that the patterns of change for law reports are generally more similar to scientific prose than to fiction or newspaper writing. Nevertheless, they assert that law reports have changed in ways that differ from the other three registers.

The authors note a rise in nominalisations in law reports, particularly over the last 50 years, which outpaces the increase found in science articles. As for colloquial features, law reports have generally been more receptive to such developments than newspaper articles and science articles, especially regarding semi-modals. In relation to

clausal complexity features, law reports have evolved in the opposite direction to the other registers considered, with a rise in nearly all types of relative clauses and noun complement clauses. As regards phrasal complexity features, law reports tend to share the upward trend discernible in the other registers.

Biber and Gray conclude that law reports “differ from most academic research writing in that they include extensive discussion of the author’s (i.e., the judge) own personal opinions and reasoning, and as a result, they show a greater receptiveness to some colloquial innovations than academic research writing” (167). Overall, however, law reports are relatively resistant to colloquial innovations.

In her chapter “Interpersonality in legal written discourse: A diachronic analysis of personal pronouns in law reports, 1535 to present” Paula Rodríguez-Puente also bases her analysis on CHELAR (Rodríguez-Puente *et al.* 2016). She argues that law reports are ‘hybrid’ in nature, being “operative in the sense that they contain a judgment or order that constitutes the actual disposition of the case, but they are also expository, since they discuss legal issues, normative facts and prescriptive legislation” (174).

Predictably, the highest frequencies are found for third person pronouns, followed by first person, with the lowest frequencies for second person, the assumption being that “legal writing is not expected to present features of personal involvement, interpersonality and subjectivity” (179). Rodríguez-Puente observes that third person narration is common in summarising the facts of the case, the singular genitive being particularly frequent. The author points out that the vast majority of pronouns are masculine rather than feminine and that such a predominance “is not a feature specific to legal writing and reflects traditional sex-role biases” (181).

First person pronouns mostly occur in judgments, frequently co-occurring with mental verbs, and are thus “the means by which authors assert their claim to speak as an authority” (186) by projecting their professional identity. Vice versa, second person pronouns are often adopted in the direct transcription of witnesses’ testimonies, in direct question-answer exchanges, or in reproducing the judge’s exact words when addressing the parties. However, the author observes that first and second person pronouns in law reports are represented “in a significantly different way from parliamentary acts, proclamations and statutes” (190).

Seen from a diachronic perspective, while the (low) frequency of second person pronouns remains constant through time, there is a marked rise in the frequency of first person pronouns, whereas third person pronouns peak in the eighteenth century, after which there is a decline, an indication that law reports “have evolved towards becoming more involved, interpersonal and subjective over time” (192).

The chapter by Nicholas Groom and Jack Grieve provides a corpus-based analysis of 130 British patent specification texts ranging from 1711 to 1860. The authors investigate how the move structure changes over time, advocating “an *evolutionary* interpretation of diachronic changes in the move structure of the patent specification genre” (205). They argue that while genre change tends to happen almost imperceptibly from the user’s perspective, it does not explain why change occurs, nor does it account for cases where genre change may even take place very rapidly. The authors suggest that an analogical evolutionary approach –which entails drawing parallels with Darwinian theory– can “easily accommodate the observation of rapid as well as gradual evolutionary change” (206), and it also explains why change occurs. Recognising the major shortcoming of this analogy with biology, Groom and Grieve propose a generalised model premised on the theory that “any process of natural or cultural change can be described as evolutionary if it exhibits three essential properties: variation, replication, and selection” (207).

The authors attempt to provide a clearly defined set of move descriptions, choosing Nasmith’s patent of 1711 (the first ever British patent) as their illustrative example. Interestingly, they find that five of the six moves in that text “appear consistently throughout the period, and thus appear in effect to constitute obligatory moves for the patent specification genre for the period under study” (214). They also note an abrupt change occurring in 1852 with new moves appearing while others disappeared, due to the passing of the Patent Law Amendment Act that year.

Groom and Grieve observe that the core moves identified “almost always occur in the same sequence order” (220). However, they also note that their “sequence data exhibit a high level of variability for most of the time period of the analysis” (220).

Anu Lehto’s chapter investigates how citizens and the British monarchy were represented in Acts of Parliament in the nineteenth and twentieth centuries. Using the *Corpus of Late Modern English Statutes*, she analyses collocates associated with these two social groups, focusing on their semantic preferences and semantic prosodies. Lehto

observes that while the Crown is usually addressed respectfully, the role of citizens evolves over time, with acts in the nineteenth century tending to focus on their criminal actions, whereas citizens' rights tend to be highlighted in the twentieth century.

The headword *person* was used to investigate the role of citizens, whereas the role of the monarch was examined by using the headwords *king*, *queen*, *crown*, *sovereign* and *monarch*. Analysing the headword *person*, the author observes that in the nineteenth century the most frequent collocates focused on general legal actions and crime, while in the twentieth century attention shifted to citizens' wellbeing and especially to children and legal documents. As for the Crown, in the nineteenth century "the topics of legal language/processes, legal actors, references to the king or queen and praise prevail, while praise, legal language/processes, and legal actors and actions are most common in the latter century" (252).

The author affirms that developments in the semantic preference and semantic prosody surrounding the headword *person* "reflect the changes made in the status of the social groups through legislation, that is, many acts were enacted in the twentieth century that specifically enhanced the rights of children and women" (255). As for the Crown, Lehto observes that the "authority of the monarch is emphasised, and diachronically the Crown is a much more central figure in the nineteenth-century acts than in the later documents" (256), with formulaic constructions appearing frequently.

In the final chapter, Claudia Claridge investigates how drunkenness was presented between 1720 and 1913 during criminal proceedings in courtroom speech by examining words meaning 'drunk'. The focus is firstly on "who uses drunkenness expressions and who is the referent of the expression" (262), and secondly on the manner of expression, "with attention paid to the forcefulness of expression" (262).

The author affirms that although "alcohol remained a constant fact of life throughout the period, attitudes toward it changed considerably" (263), with a less tolerant view towards it appearing during the Victorian period.

Given that the texts were taken down in shorthand by scribes in the courtroom, the author points out that while they present a reasonably accurate picture of historical speech, "they are nevertheless not an ideal representation of forensic interaction" (265).

The author observes that male voices clearly predominate, accounting for 20 million of the 24 million words. Of the 30 or so expressions relating to drunkenness

found in the corpus, the two most frequently used expressions are *drunk* and *in liquor*: a variety of sometimes colourful terms “also indicate various strength of intoxication” (272). She asserts that “the majority of drunkenness references remains unmodified (67.2%)” (280), while amplification is found in 24.7 per cent of cases and downtoning in 8.1 per cent of cases.

Claridge concludes that, with regard to drinking, judges and defendants seem to have conflicting agendas. While defendants keep on using incapacitation through drink as a potential mitigating factor, the court’s insistent probing in questions is apparently more in the interest of laying blame (on defendants as well as victims) than on finding excuses. The defendants’ misconception of (evolving) legal views is also apparent in their willingness to attribute intoxication to themselves and to even amplify it (283).

A considerable number of specific and general corpora have been used by the authors in this volume in carrying out their research, a reflection of the “corpus-based” nature of the volume itself. The law-related corpora and databases used by the authors are the following:

- the *British Hansard Corpus* (Lastres-López);
- the parallel (English and French) *Canadian Hansard Corpus* (Lastres-López);
- the O.J Simpson trial (see Linder 2017) (Reppen and Chen);
- the *British Law Report Corpus* (see Marín Pérez and Rea Rizzo 2012) (Goźdz-Roszkowski);
- the *American Law Corpus* (Goźdz-Roszkowski);
- CHELAR (see Fanego *et al.* 2017) (Biber and Gray; Rodríguez-Puente);
- the *Corpus of Early Modern English Statutes 1491–1707* (Rodríguez-Puente);
- the historical archive of British patents held at the British Library (Groom and Grieve);
- the *Corpus of Late Modern English Statutes* (Lehto);
- the *Old Bailey Corpus* (Claridge).

The following generic reference corpora were also adopted:

- the *British National Corpus* (Breeze; Reppen and Chen);

- the British component of the *International Corpus of English* (see Nelson *et al.* 2002) (Lastres-López);
- the French and Spanish subcorpora of the *Integrated Reference Corpora for Spoken Romance Languages* (see Cresti and Moneglia 2005) (Lastres-López);
- the *Corpus of English Dialogues* (see Kytö and Walker 2006) (Reppen and Chen);
- the *Helsinki Corpus of English Texts* (see Rissanen *et al.* 1991) (Rodríguez-Puente).

One author (Diani) relied entirely on the corpora she had compiled herself, while in four cases the authors used a combination of corpora together with ones they had compiled themselves (Breeze, Lastres-López, Goźdz-Roszkowski, Biber and Gray).

## 2. DISCUSSION

Over the last 20 years or so, the exponential growth of corpora and databases in the legal sphere has radically transformed the way people approach law-related matters, opening up new horizons that are both exciting and overawing. Accessing enacted legislation and court judgments online has become routine for today's legal professionals. In the academic sphere, this digital revolution has produced a rich variety of corpora on legal discourse, both synchronic and diachronic, in many cases compiled by linguists (for an overview up to 2012, see Pontrandolfo 2012). Notable instances of legal corpora have been created, for example, in the United States, the United Kingdom, Spain, Italy, Germany and Poland. The linguistic study of legal corpora is sometimes referred to as 'legal corpus linguistics' (see, e.g., Hamann and Vogel 2017: 101; Dale 2018). Computer-Assisted Legal Linguistics (CAL) is a new subfield denoting the research carried out by the interdisciplinary group whose members explore "the fabric of language and law" (Hamann and Vogel 2017: 101). Law and Corpus Linguistics (LCL) tends to refer above all to the research carried out by scholars involved in the Law and Corpus Linguistics conferences held at Brigham Young University (several of whom are also part of the CAL project) where Mark Davies and his team have compiled numerous large corpora, both generic and specialised.

In the United States, judges and scholars are debating whether corpus linguistics should be used as a “tool in legal interpretation” (Solan and Gales 2017: 1311) in court cases, for example when the “ordinary meaning” (Lee and Mouritsen 2018: 788) of a word or expression needs to be established. The first case of corpus linguistics being used in the US Supreme Court dates back to 2011 (Zimmer 2011). However, in ways that are slightly analogous to the ongoing debate about whether plain language should be used in legally binding texts (see Williams 2011), there are also judges and scholars who warn against the pitfalls of using corpus linguistics in court cases (see, e.g., Ehrett 2019; United States Court of Appeals for the Sixth Circuit 2019: 23-26).

In academia, on the other hand, corpora are widely accepted as a valuable tool for analysing linguistic phenomena in general, including legal discourse. Indeed, using corpora has become a mainstream activity among linguists rather than a niche area as was the case, say, thirty years ago. Besides (or instead of) making use of the corpora available online, many researchers create their own corpora which can be tailor-made to highlight the particular aspect of research they wish to investigate. Scholars of legal corpora are generally interested in exploring one or more of the following fields: legal translation and interpreting, the teaching of law-related matters (particularly legal language), and the linguistic features of legal discourse.

*CRVELD* comes within the latter strand of research, and is a very welcome addition to the field for reasons I will shortly outline. It is no coincidence that the two co-editors themselves were part of the team that compiled CHELAR. Under the guidance of Teresa Fanego, scholars belonging to the *Research Unit for Variation, Linguistic Change and Grammaticalization*, originally set up in the 1990s at the University of Santiago de Compostela, have predominantly focused on diachronic linguistic phenomena, and in recent years this has included research on legal discourse. The Research Unit is also part of an international consortium of universities involved in expanding and tagging ARCHER, the diachronic corpus of British and American English registers originally compiled by Douglas Biber and Edward Finegan in the early 1990s.

Although corpus-based research on legal discourse has so far spawned little more than a handful of monographic works (e.g., Archer 2005; Heffer 2005; Torikai 2006; Goźdz-Roszkowski 2011; Kopaczyk 2013; Larner 2014; Lehto 2015), it has given rise to a vast array of articles and book chapters, many appearing in academic journals, or in

edited volumes either on legal discourse or on corpus studies. That said, there are relatively few edited volumes devoted to corpus-based studies on legal discourse. Exceptions include Mori (2018) where eleven EU languages ('Eurolects') are analysed, Goźdz-Roszkowski and Pontrandolfo (2017) which focuses on phraseology in a variety of legal and institutional settings, and the special issue of *Brigham Young University Law Review* on law and corpus linguistics (2017) which essentially examines whether and how corpora can be used in legal interpretation, particularly in the USA but also in Germany.

As we have seen, *CRVELD* focuses on two theme areas: 'Cross-genre and cross-linguistic variation' and 'Diachronic variation'. The first half, therefore, has some elements in common with the volume edited by Goźdz-Roszkowski and Pontrandolfo (2017), notably the two chapters with a phraseological orientation, respectively by Reppen and Chen and Goźdz-Roszkowski. However, given the key role of analysing phraseology (lexical bundles) in corpus linguistics in general, and the vastness of the universe known as legal discourse, there is room for a considerable number of studies on the subject without the risk of redundancy. As for the other three chapters in Part One of *CRVELD*, respectively by Diani, Lastres-López, and Breeze, the first two propose a cross-linguistic analysis, which again is one of the major themes in Goźdz-Roszkowski and Pontrandolfo (2017) though seen here from a different (i.e. non-phraseological) perspective, while Breeze examines parts-of-speech patterns.

Moreover, Part One of *CRVELD* displays internal coherence in terms of adhering to two of Biel's (2010) four trajectories, the first two chapters corresponding to Trajectory 4 (cross-linguistic variation), while the latter three come within Trajectory 2 (internal variation).

Where *CRVELD* diverges most markedly from other edited volumes on legal corpus linguistics to date is in the five chapters devoted to diachronic variation, the area the two co-editors have explored extensively in their academic research. In my opinion, two of these chapters in particular deserve further comment, respectively the ones by Biber and Gray and Groom and Grieve. Both chapters are dense in terms of the information and analyses they present and cannot be read hurriedly. However, the effort involved is rewarded by the fact that in both cases the authors manage to go beyond the mere presentation of data (a weakness, it has to be said, of some corpus-based research, though not in this particular volume) and contextualise the results in ways that are

original and thought-provoking. In the former case, this is done by affording a nuanced view as to exactly how law reports may differ over time, very often (but not always) by resisting change, with respect to other genres. In the latter case, by applying evolutionary theory to the study of genre change, the authors provide an engaging narrative by highlighting in detail the ways in which the genre of patents adapted to changing historical circumstances between 1711 and 1860.

This is in no way to belittle the other contributions in this volume, all of which provide intriguing and original analyses of the topics selected. The chapters by Breeze, Goźdz-Roszkowski, and Rodríguez-Puente, for example, all stand out as carefully thought-out pieces of scholarship on highly technical issues. Although several of the topics in the volume are indeed quite complex, they are all presented in a coherent and attractive way so that the reader can follow the line of reasoning of each chapter without undue difficulty. As Xin and Wang (2019: 131-132) rightly point out, the readability of *CRVELD* “is enhanced with clearly illustrated tables, diagrams and figures, as well as with updates on technical innovations in corpus linguistics.”

Taken as a whole, then, *CRVELD* represents a step forward in terms of our knowledge of legal corpus linguistics. This applies not only to the insights offered by scholars in Part One on cross-genre and cross-linguistic variation but in particular to Part Two of the volume devoted to diachronic variation. Of course, individual articles and papers on the historical dimension of legal corpus linguistics are available in numerous journals and edited volumes, but having five chapters together on the topic gives added depth to this area of study. In this respect I disagree with Xin and Wang (2019: 132) when lamenting the volume’s “lack of phonetic, phonological, multimodal or multilingual analysis.” In my view, restricting the range of topics of an edited volume endows it with greater internal coherence. Indeed, I would hope that future edited volumes on legal corpus linguistics will focus on increasingly limited areas of this rapidly evolving field of inquiry.

All of the contributors to *CRVELD* are linguists, ranging from scholars of international renown to postdoctoral researchers, thus representing a heterogeneous mixture, also in terms of geographical provenance and linguistic backgrounds. However, also in the light of the recent interest of judges, law scholars and legal experts in using corpora in the courts in the world’s major English-speaking nation, as is highlighted in *Brigham Young University Law Review* (2017) which contains papers

from linguists and law scholars in roughly equal measure, another suggestion for future edited volumes on legal corpus linguistics might be to make them truly interdisciplinary by involving linguists (including computational linguists), law scholars and legal professionals, among others, so as to offer a plurality of insights into the same theme area.

While *CRVELD* will be of particular interest to scholars of legal corpus linguistics, in my view it has the requisites to arouse the curiosity of corpus linguists approaching legal discourse for the first time, and also of legal scholars who have so far shied away from corpora. Rather than attempting to read the book from cover to cover (how often do we actually complete that endeavour these days, especially with edited volumes?), I would suggest starting with the excellent introductory chapter by Teresa Fanego and Paula Rodríguez-Puente and then choosing a chapter (or two) that appeals most to one's individual taste and research interests, and taking it from there. Readers will not be disappointed.

#### REFERENCES

- Archer, Dawn. 2005. *Questions and Answers in the English Courtroom (1640–1760): A Sociopragmatic Analysis*. Amsterdam: John Benjamins.
- Bhatia, Vijay K. 1987. Language of the law. *Language Teaching* 20/4: 227–234.
- Bhatia, Vijay K. 1993. *Analysing Genre. Language Use in Professional Settings*. Harlow: Pearson Education.
- Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.
- Biel, Łucja. 2010. Corpus-based studies of legal language for translation purposes: Methodological and practical potential. In Carmen Heine and Jan Engberg eds. *Reconceptualizing LSP. Online Proceedings of the XVII European LSP Symposium 2009*. Aarhus: Aarhus School of Business, Aarhus University.
- Brigham Young University Law Review. 2017. Special issue on law and corpus linguistics (vol. 6, 2017). <https://digitalcommons.law.byu.edu/lawreview/vol2017/iss6/>
- Cresti, Emanuela and Massimo Moneglia eds. 2005. *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: John Benjamins.
- Culpeper, Jonathan and Merja Kytö. 2010. *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge: Cambridge University Press.
- Dale, Kezziah. 2018. Legal corpus linguistics: *Gambling to gaming* language powers and probabilities. *UNVL Gaming Law Journal* 8/2: 233–252.

- Ehrett, John S. 2019. Against Corpus Linguistics. *The Georgetown Law Journal Online* 108: 50–73.
- Fanego, Teresa, Paula Rodríguez-Puente, María José López-Couso, Belén Méndez-Naya, Paloma Núñez-Pertejo, Cristina Blanco-García and Iván Tamaredo. 2017. The *Corpus of Historical English Law Reports 1535–1999 (CHELAR)*: A resource for analysing the development of English legal discourse. *ICAME Journal* 41: 53–82.
- Goźdz-Roszkowski, Stanislaw. 2011. *Patterns of Linguistic Variation in American Legal English: A Corpus-Based Study*. Bern: Peter Lang.
- Goźdz-Roszkowski, Stanislaw and Gianluca Pontrandolfo eds. 2017. *Phraseology in Legal and Institutional Settings: A Corpus-based Interdisciplinary Perspective*. London: Routledge.
- Hamann, Hanjo and Friedemann Vogel. 2017. The fabric of language and law – Towards an international research network for Computer Assisted Legal Linguistics (CAL). *International Journal of Language & Law* 6: 101–109.
- Heffer, Chris. 2005. *The Language of Jury Trial: A Corpus-aided Analysis of Legal-lay Discourse*. Basingstoke: Palgrave Macmillan.
- Hundt, Marianne and Christian Mair. 1999. “Agile” and “uptight” genres: The corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4/2: 221–242.
- Kopaczky, Joanna. 2013. *The Legal Language of Scottish Burghs: Standardization and Lexical Bundles (1380-1560)*. Oxford: Oxford University Press.
- Kytö, Merja and Terry Walker. 2006. *Guide to A Corpus of English Dialogues 1560–1760*. Uppsala: Acta Universitatis Upsaliensis.
- Larner, Samuel. 2014. *Forensic Authorship Analysis and the Word Wide Web*. Basingstoke: Palgrave.
- Lee, Thomas R. and Stephen C. Mouritsen. 2018. Judging ordinary meaning. *Yale Law Review* 127/4: 788–879.
- Lehto, Ana. 2015. *The Genre of Early Modern English Statutes: Complexity in Historical Legal Language*. Helsinki: University of Helsinki dissertation.
- Linder, Douglas O. 2017. Famous Trials. <http://www.famous-trials.com/> (1 February, 2017.)
- Marín Pérez, María José and Camino Rea Rizzo. 2012. Structure and design of the *British Law Report Corpus (BLRC)*: A legal corpus of judicial decisions from the UK. *Journal of English Studies* 10: 131–145.
- Mori, Laura ed. 2018. *Observing Eurolects: Corpus Analysis of Linguistic Variation in EU Law*. Amsterdam: John Benjamins.
- Nelson, Gerald, Sean Wallis and Bas Aarts. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.
- Pontrandolfo, Gianluca. 2012. Legal Corpora: An overview. In Marella Magris and Helena Lozano eds. *Rivista Internazionale di Tecnica per la Traduzione* 14: 121–136.
- Rissanen, Matti, Merja Kytö, Leena Kahlas-Tarka, Matti Kilpio, Saara Nevanlinna, Irma Taavitsainen, Terttu Nevalainen and Helena Raumolin-Brunberg. 1991. *The Helsinki Corpus of English Texts*. <http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/>
- Rodríguez-Puente, Paula, Teresa Fanego, María José López-Couso, Belén Méndez-Naya and Paloma Núñez-Pertejo, Paloma. 2016. *Corpus of Historical English*

Law Reports 1535–1999 (CHELAR).  
<http://www.helsinki.fi/varieng/CoRD/corpora/CHELAR/>.

- Šarčević, Susan. 2000. *New Approach to Legal Translation*. The Hague: Kluwer Law International.
- Solan, Lawrence and Tammy Gales. 2017. Corpus Linguistics as a tool in legal interpretation. *Brigham Young University Law Review*: 1311–1358.
- Swales, John M. 1990. *Genre Analysis: English for Academic and Research Settings*. Cambridge: Cambridge University Press.
- Tiersma, Peter M. 1999. *Legal Language*. Chicago IL: The University of Chicago Press.
- Torikai, Shinichiro. 2006. *A Corpus-based Study of Legal English: Investigating the Language of the House of Lords Judgments 1677–2000, with Particular Reference to Reported Discourse*. Lancaster: University of Lancaster dissertation.
- United States Court of Appeals for the Sixth Circuit. 2019. *Wilson v. Safelite Group, Inc.* <http://www.clarkcunningham.org/JP/WilsonVSafelite-6thCir.pdf>.
- Williams, Christopher. 2011. Legal English and plain language: An update. *ESP Across Cultures* 8: 139–151.
- Xin, Zhiying and Jiawei Wang. 2019. Review of *Corpus-based Research on Variation in English Legal Discourse*, Teresa Fanego and Paula Rodríguez-Puente, eds (2019). *The International Journal of Speech, Language and the Law* 26/1: 127–132.
- Zimmer, Ben 2011. The corpus in the Court: ‘Like lexis on steroids’. *The Atlantic*. <https://www.theatlantic.com/national/archive/2011/03/the-corpus-in-the-court-like-lexis-on-steroids/72054/> (4 March, 2011.)

*Reviewed by*  
 Christopher Williams  
 e-mail: [christopher.williams@unifg.it](mailto:christopher.williams@unifg.it)

Review of Doval, Irene and María Teresa Sánchez Nieto eds. 2019. *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications*. Amsterdam: John Benjamins. ISBN: 978-9-027-20234-5. <https://doi.org/10.1075/scl.90>

Roberto A. Valdeón  
University of Oviedo / Spain and Jinan University Zhuhai / China

In *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications*, published in the prestigious John Benjamins' *Studies in Corpus Linguistics* series, Irene Doval and María Teresa Sánchez Nieto have gathered a selection of the contributions to the International Conference *Parallel Corpora: Creation and Applications*. The conference, held at the University of Santiago de Compostela (Spain) in 2016, focused on the exploitation of parallel corpora for diverse purposes, and more precisely for contrastive and translation studies. As the editors posit in their introduction, since the 1990s the use of corpora has changed the ways in which language and language in practice have been studied, as comparable and parallel corpora have served researchers to investigate differences and similarities between languages. Some of the first corpora (such as the *English-Norwegian Parallel Corpus* and the *English-Swedish Parallel Corpus*) had a clear academic purpose. Others have functioned as language resources widely used by researchers even though they were not the result of an academic endeavour as such, e.g. the multilingual corpora of the various European Union institutions. As can be expected, these contain institutional language and, hence, can be used for specific research purposes. But the potential of these resources, and of corpora in general, has kept growing over the past two decades. As Doval and Sánchez Nieto (3) remind us, parallel and comparable corpora are now used in machine translation and multilingual natural language processing, contrastive studies, translatology, lexicography, and also the teaching of foreign language and translation.

In the case of translation studies, Doval and Sánchez Nieto add, corpora have been particularly useful in applying a more empirical paradigm to descriptive studies, and, one would hope, to go beyond the many limitations of descriptivism. Many of the challenges discussed by the contributors to *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications* have been recently highlighted by De Sutter and Lefer (2020: 18–19) who, in an article on a new agenda for corpus-based translation studies, have argued for

a new multifactorial, multi-methodological and interdisciplinary research agenda for empirical translation studies [...] that can potentially help us to characterize translated text, starting with linguistic features that have been said to typify other forms of constrained communication, such as non-native language varieties, editing and student writing,

moving the focus away from non-crucial parts of the corpus-based research agenda such as the study of universals (2020: 2). In addition, De Sutter and Lefer discuss corpora as process and product.

Indeed, to conclude their introduction, the editors of the book stress the two main trends in today's parallel corpora, i.e. the design and building of corpora on the one hand, and the features and applicability of corpora as products on the other. This distinction will serve to guide the readers through the well-structured contents of the collection. It is surprising, though, that no working definition of the central concept discussed in the book is provided. It is true that, although the literature abounds with definitions, it might be a mission impossible to find one that is widely used: see, for example, definitions in Olohan (2004: 24–25, 35–37) or Mikhailov and Cooper (2016: 2–8), especially in a collection of articles by various authors who are likely to use the term in slightly different ways. However, a reference to this crucial problem might serve as word of caution for the non-specialist yet interested reader.

The book is divided into three uneven sections, namely 1) background and processing, 2) creation, annotation and access, and 3) tools and application. The first section starts with a valuable introductory article on the name and nature of comparable parallel corpora. Hareide discusses a couple of definitions before providing her own “two or more parallel/translation-corpora that have the same sampling” (21), in order to underscore one of the main problems with many corpus-based translation studies, i.e. their replicability. To illustrate the usefulness of working with larger corpora defined by a number of specific parameters, Hareide uses the *Norwegian Spanish Parallel Corpus*

and the *English-Spanish P-ACTRES* to test Sandra Halverson's so-called *Gravitational Pull Hypothesis*. Hareide, who tests grammatical structures, posits that the results confirm the need of "well-designed and correctly used corpora" (34).

In the next three chapters of this section, Josep Marco, Rosa Rabadán and Martin Volk offer somehow alternative uses and applications of parallel corpora. Marco defends the use of parallel corpora for two purposes: as the main source of information to analyse translators' choices and as secondary data to supplement information provided by a comparable corpus. Here we have the first difference with Hareide's combined use of 'comparable' and 'parallel' and, hence, the first example of potential confusion for the non-specialist even though, ultimately, both Hareide and Marco's efforts go in the same direction. Marco uses two examples to show the value of parallel corpora. Both draw on the *Valencia Corpus of Translated Literature* or *COVALT* to find patterns of correspondence between source and target texts when using parallel corpora, and to explain certain patterns when using comparable corpora to complement the former.

In line with Marco's chapter, Rabadán insists on the importance of combining parallel corpora with comparable and monolingual corpora. Starting with broad and narrow definitions of parallel corpora, which do not correspond exactly to those of previous chapters, Rabadán defends the need to recycle or reuse existing corpora rather than to waste time to build a new one, even if that involves upgrading existing sources to meet the demands of new research projects. Corpus efficiency can also be achieved by using comparable and monolingual corpora, or by adding new annotations to the information stored in those resources. Rabadán also includes a set of useful strategies to enhance collaborative efforts when embarking on a new project and, thus, save precious time (71). It is also worth noting that in the second section of the book, Doval *et al.* will argue in favour of creating new corpora when existing ones do not provide relevant information for specific research questions, and Sanjurjo-González and Izquierdo will claim that the creation of the *P-ACTRES Parallel Corpus* at the University of León was precisely the result of the insufficient nature of the *Cobuild* and *CREA* corpora that Rabadán and her colleagues had used until then.

For his part, Martin Volk draws on his own experience in gathering a variety of Swiss databases to emphasise the need of appropriate word alignment in parallel corpora in order to improve annotation, which in turn would be beneficial for more

practical purposes, such as language learning and computational linguists wanting to evaluate the quality of automatic sentence alignment. For this purpose, standard annotations methods such as Part of Speech tagging, he claims, should be complemented with language-specific methods to deal with, for example, split verbs in German. Volk also presents prototypes that can improve word alignment and annotation and, therefore, contribute to dealing with issues such as translation error detection.

Section II comprises a total of nine chapters that delve into corpora creation, annotation and access. Some authors build upon corpora already discussed in the previous section (Molés-Cases and Oster, Sanjurjo-González and Izquierdo), while the rest introduce new ones, ranging from the intermodal corpus of *European Parliament Speeches* or EPTIC (Ferraresi and Bernardini) to the smaller *Corpus of German-Basque Literary Translations* (Sanz-Villar). In addition, some authors discuss corpora built from scratch (Doval *et al.*), while others present spin-offs from larger (e.g. Čermák on *InterCorp*, part of the *Czech National Corpus*) or different corpora (Molés-Cases and Oster on *COVALT PAR\_ES*).

The section describes a number of important issues as regards creating and annotating corpora, which highlights the specific requirements of the various databases discussed. The contributors provide information on the features and challenges faced by the creators of these corpora as well as on the annotation processes. Of particular note is the article by Čermák, who underscores the amount of work required to build up a specific corpus, i.e. *InterCorp*, based at Charles University in Prague. *InterCorp* comprises texts in Czech and in other thirty-nine languages, aligned by a team of nearly 200 individuals (85). The chapters also show that the larger corpora, and subcorpora, tend to include at least one widely spoken language (typically English, French, German and Spanish), whereas lesser spoken languages (e.g. Vietnamese or Catalan in *InterCorp* (96–97), are more likely to be found in smaller corpora or subcorpora, except when the database specialises on a specific language: for example, Galician in the *CLUVI* Corpus (Gómez Guinovart), Catalan in *COVALT* (Marco) or Finnish in *PEST* (Mikhailov *et al.*). The case of *TAligner* is particularly interesting as its compilation includes German and Basque but also Spanish, as many literary texts were translated indirectly from Spanish (Sanz-Villar). Some chapters provide an innovative approach by including multimodal texts (Doval *et al.*) or what is termed as an intermodal corpus (Ferraresi and Bernardini).

As regards use, the various corpora presented in this section allow different possibilities. *InterCorp* allows users to compare up to four languages and to search one or more languages by phrase or by lemma (Čermák), *PaGeS* (Doval *et. al.*) and *PEST* (Mikhailov *et al.*) to compare dialectal variation (Doval *et. al.*), *EPTIC* to compare different communication modes (Ferraresi and Bernardini), *MULTINOT* to study contrastive differences between original texts in English and Spanish, between translations in both directions and between translated *versus* non-translated texts in both languages (Lavid López). Some have been used to produce bilingual dictionaries (e.g. *InterCorp*) or might improve computational systems in different subfields in the future (e.g. *MULTINOT*). Most authors stress the dynamic nature of these corpora, which allows them to set up specific goals to do research at present while also providing an opportunity to consider different objectives as the corpora evolve.

Finally, the three chapters in Section 3 cover the tools and applications of comparable and parallel corpora. Pablo Gamallo Otero discusses techniques to build highly reliable bilingual dictionaries using comparable corpora to test the validity of the choices made when creating a new dictionary by using two existing ones, and shows their value to create new dictionaries for languages with fewer resources and parallel corpora. García *et al.* also draw on Iberian languages, i.e. Spanish and Portuguese, to suggest the use of parallel corpora to extract bilingual collocation equivalents. Given the tendency by non-native speakers of a language to use unusual lexical combinations, García *et al.* stress that parallel corpora can be used to identify thousands of collocation equivalents with a very high precision (of around 86%) in an automatic and fast manner. This would allow the production of dictionaries and other teaching materials for language classroom use. It is true, however, that the success of the strategy would need to be tested with less closely related languages in order to confirm its usefulness with other language pairs. Finally, Ghoshal and Rao explore normalization processes of abbreviations and shorthand forms in French text messages. Although their experiment was successful, the objective of their work is never clearly explicated.

On the whole, the chapters in this collection make a strong case for the use of parallel, comparable, bidirectional (Lavid López; Sanjurjo-González and Izquierdo), multilingual (Čermák), intermodal quasi-parallel (Ferraresi and Bernardini) and comparable parallel (Hareide) corpora in contrastive and translation studies. They underscore their potential not only for descriptive studies but also for translator training,

translation practice, machine translation, post-editing dictionary making and so on. Most of them focus on linguistic aspects (e.g. motion events in German and Spanish, adverbials in English and Catalan, Spanish gerunds and the corresponding forms in Norwegian and English, modality) that could be examined by means of corpora. But the chapters may also provide ammunition to those translation scholars who, according to Malmkjaer (1998) - quoted by Marco in this volume - feel a “disaffection bordering on hostility [...] with regard to linguistics” (43). Hopefully, the arguments carefully laid out by the authors and the editors of this volume will entice some of the disaffected to the dark side as well.

#### REFERENCES

- De Sutter, Gert and Marie-Aude Lefer. 2020. On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach. *Perspectives* 28/1: 1–23.
- Malmkjaer, Kirsten. 1998. Love thy neighbour: Will parallel corpora endear linguists to translators? *Meta: Translator's Journal* 43/4: 534–541.
- Mikhailov, Mikhail and Robert Cooper. 2016. *Corpus Linguistics for Translation and Contrastive Studies*. London: Routledge.
- Olohan, Maeve. 2004. *Introducing Corpora in Translation Studies*. London: Routledge.

*Reviewed by*  
 Roberto A. Valdeón  
 University of Oviedo  
 Department of English, French and German.  
 C/ Amparo Pedregal s/n  
 E-33011, Oviedo.  
 Spain  
 e-mail: [valdeon@uniovi.es](mailto:valdeon@uniovi.es)

Review of Amador Moreno, Carolina P. 2019. *Orality in Written Texts: Using Historical Corpora to Investigate Irish English (1700–1900)*. London: Routledge. ISBN: 978-1-138-80234-6. <https://doi.org/10.4324/9781315754321>

Raymond Hickey  
University of Duisburg and Essen / Germany

The use of alternative data sources in historical linguistics has become a common procedure in recent years, moving away from literary texts as documentation for language variation and change. The current book is an example of this approach, in this case availing of personal letters for insights into language use in previous centuries, specifically the use of English on the island of Ireland in its different forms, in the north and south of the country. In this respect it is in the company of such volumes as Fitzmaurice (2004), Dossena and Del Lungo Camiciotti (2012), Auer *et al.* (2015), Hickey (2019) on letters and in the wider context of ego-documents (van der Wal and Rutten 2013).

This monograph consists of 203 pages of primary text, divided into seven sections: the first is an introduction (1–35) and the last some concluding remarks (200–203). In between there are five chapters. The first deals with the historical contexts of letters, the second with the nature of orality in such private correspondence while the other three chapters are dedicated to the linguistic analysis of discourse-pragmatic variation, of deixis and of embedded questions in the letters.

The corpus of correspondence, which forms the data base for the author's investigation, is the *Corpus of Irish English Correspondence* (CORIECOR) which she has been compiling with her Irish colleague Kevin McCafferty from the University of Bergen over the past decade (see positioning article, McCafferty and Amador-Moreno 2012).

The introduction provides the context for the book as a whole, reporting on and discussing literature on historical sociolinguistics and the use of ‘bad data’ (fragmentary but relevant data for a topic). The author also provides background information on the study of Irish English and on historical corpora and their applicability when engaging in variationist studies. There is also detailed information on her own corpus and a discussion of the challenges it presents for linguistic analysis.

Chapter two goes into more detail concerning the historical context of letters, who the writers were and who the recipients were and what the motivation for emigration was in eighteenth- and nineteenth-century Ireland. In addition, the question of literacy is broached. This is of central concern in the Irish context as the greatest level of literacy was to be found among northern Protestants. Hence this group tends to be over-represented in the CORIECOR corpus and is a factor which must be borne in mind if any generalisations are to be made from the emigrant correspondence to diachronic Irish English in the two centuries covered by the CORIECOR corpus.

The linguistic core of the present book begins with Chapter 3 “The orality of private correspondence. Using emigrant letters for linguistic analysis.” The discussion here is concerned with how reliable written data is for reconstructing what the author calls ‘the voices of the writers’. In particular, the question arises whether all instances of irregular orthography represented an underlying phonetic reality for the letter writer. The use of *the* for *they* is just once instance, discussed by the author. Furthermore, there is the issue of formulaic language. The tendency for this to occur is greatest in letter openings and closings and so does not impinge on the body of a letter, but must be borne in mind, nonetheless.

Chapter 4 launches straight into an analysis of discourse pragmatic markers, first discussing Seamus Heaney’s use of *so* at the beginning of his translation of *Beowulf* but quickly moving to the CORIECOR data where the author investigates how *so* is employed by letter writers. Other markers are treated in further sections of the chapter, e.g. *anyhow/anyway*, *like*, *sure*, the latter two being particularly relevant to the pragmatics of many varieties of English today, not just Irish English. The documentation of findings is meticulous in this chapter with detailed statistics of occurrences in the CORIECOR corpus.

Chapter 5 is dedicated to a detailed consideration of deixis in the letter corpus. This term is understood broadly by the author with forms which have particular usages

in Irish English being given special attention, e.g. *there*. The chapter also considers the occurrences of personal pronouns, both in the letter corpus and in the author's own Irish-Argentine sub-corpus which provided interesting corroborations of findings elsewhere.

Chapter 6 is the third and last of the data investigation chapters and is dedicated to examining the occurrence of embedded inversion in the letters considered. By 'embedded inversion' is meant the use of question word order in verbal complements, e.g. *We wondered was he coming home*. The author considers the occurrence of this word order across many varieties of English and then looks at the evidence presented by her correspondence corpus which shows a definite preference for embedded inversion. She is aware of the fact that such inversion is the rule in the Irish language and many of the correspondents would have been shifters from Irish to English or have spoken a contact variety of the latter deriving from recent language shift. The syntactic and lexical contexts in which this embedded inversion is found are scrutinised in considerable detail and comparisons with other corpora are made. The author also considers whether there are 'privileged points of entry' for non-standard features to become established in a variety, a notion close to the much discussed phenomenon of salience in language which can further the spread of features if low and hinder this when high.

In summary, one can say that this book provides a timely addition to the burgeoning field of private letter analysis for linguistic purposes. It shows clearly what insights into non-standard grammatical and pragmatic features can be gleaned from a close examination of emigrant correspondence. This can serve many purposes including that of confirming the existence of features in vernacular, quasi-oral language of the past, features which are not necessarily evident in other text types.

However, a reader coming to this book without any prior knowledge of Irish English might be forgiven for thinking that the features discussed are the only ones, or at least the most prominent in this correspondence and, by implication, in Irish English. But even the briefest contact with Irish English shows that a large number of non-standard features are to be found in verbal syntax – non-standard verbal concord, tense usage and, above all, aspectual distinctions not found in more standard forms of English (see the detailed treatment in Hickey 2007). Incidentally, these features are also reflected in the language of the CORIECOR corpus. So it would have been of great

benefit if the author had explained at the outset why she chose the particular traits which she dedicated chapters to in her book. True, she has used material from her previous publications (and acknowledges this), but in a book-length study there is room for more discussion or, at the very least, a clear indication of features which exist in the correspondence but which the author decided not to discuss, for whatever reason.

Finally one can say that this book reads well and that the author displays a wide knowledge of the topic and its framework, discussing much relevant literature and many different but comparable investigations to hers. It is well type-set and indexed and should provide both students and scholars researching varieties of English with an important source of information on historical Irish English correspondence.

#### References

- Auer, Anita, Daniel Schreier, Richard J. Watts eds. 2015. *Letter Writing and Language Change*. Cambridge: Cambridge University Press.
- Dossena, Marina and Gabriella Del Lungo Camiciotti eds. 2012. *Letter Writing in Late Modern Europe*. Amsterdam: John Benjamins.
- Fitzmaurice, Susan 2004. *The Familiar Letter in Early Modern English. A Pragmatic Approach*. Amsterdam: John Benjamins.
- Hickey, Raymond 2007. *Irish English. History and Present-day Forms*. Cambridge: Cambridge University Press.
- Hickey, Raymond ed. 2019. *Keeping in Touch. Familiar Letters Across the English-Speaking World*. Amsterdam: John Benjamins.
- McCafferty, Kevin and Carolina P. Amador-Moreno 2012. A Corpus of Irish English Correspondence (CORIECOR): A tool for studying the history and evolution of Irish English. In Bettina Migge and Máire Ní Chiosáin eds. *New Perspectives on Irish English*. Amsterdam: John Benjamins, 265–288.
- Van der Wal, Marijke J. and Gijsbert Rutten 2013. *Touching the Past: Studies in the Historical Sociolinguistics of Ego-Documents*. Amsterdam: John Benjamins.

*Reviewed by*

Raymond Hickey

Universitätsstr.12

D-451141 Essen

Germany

e-mail: [raymond.hickey@uni-due.de](mailto:raymond.hickey@uni-due.de)