RiCL

Research in Corpus Linguistics

RiCL 11/1 (2023)

aelinco
Asociación Española de Lingüística de Corpus

RiCL 11/1 (2023)

Editors

Paula Rodríguez-Puente and Carlos Prado-Alonso

ISSN 2243-4712

https://ricl.aelinco.es/

RiCL Research in Corpus Linguistics

Official journal of aelinco Asociación Española de Lingüística de Corpus

RiCL Research in Corpus Linguistics

11/1 (2023)

RiCL Research in Corpus Linguistics

The synchronous and asynchronous learning of anaphora: A corpus-based analysis with learners of English and Spanish

Amanda Maraschin Bruscato – Jorge Baptista
University of Algarve / Portugal

Abstract – This paper aims to investigate the use of nominal, pronominal, and zero anaphora among native speakers of Brazilian learning Spanish or English. To this purpose, two learner corpora were employed: the *Brazilian Learners of Anaphora in English* (BRANEN) and the *Aprendices Brasileños de Anáfora en Español* (BRANES). Participants were undergraduate students with an intermediate-to-advanced proficiency level in the foreign language (English or Spanish) and were randomly assigned into three groups: one had synchronous lessons on the topic, one had asynchronous lessons, and a third one was the control group (which had no lesson). They all completed short narratives in four moments, and their written texts were compiled to investigate how a different instructional mode can better contribute to the learning of this specific discourse mechanism. Third-person human subjects of finite clauses and their antecedents were manually annotated on *Recogito*. When analysing the pre-test, we found that learners could be less redundant and could use more zero anaphora than pronominal anaphora in English coordinate clauses and Spanish main clauses to continue the topic/subject. The experimental groups practised it during the online course and the asynchronous instructional mode proved to be more effective until the third test (immediately after the course), but the same was not found on the delayed post-test (one month later).

Keywords – anaphora; language learning; asynchronous learning; synchronous learning; learner corpus

1. INTRODUCTION

The processing of anaphora has been the focus of many studies over the past years, due to the importance of the mechanism for a cohesive communication in any language. In simple terms, anaphora can be defined as a discourse mechanism in which an element in the text (anaphor) refers back to another element (antecedent), as in (1), where the pronouns *she* and *her* refer back to *Mary*.

(1) **Mary** fell. **She** was still on the ground when Peter found **her**.

In the previous example, there was only one possible antecedent, but in discourse there are usually more possibilities. Although the human brain is generally able to identify the correct antecedents, natural language processing has been a challenge to computational linguistics, due to the difficulty of training a machine to understand how the human brain works. Thus, many researchers have been trying to comprehend better how speakers of different languages correlate anaphors and antecedents.

As Lozano (2021a) suggests, research on anaphora resolution is relevant to investigate cross-linguistic influence and second language (L2) development. Many studies have analysed anaphora resolution through questionnaires with ambiguous sentences or through learner corpora, as onesome of the works presented in the first international conference on *The Acquisition and Processing of Reference and Anaphora Resolution* (APRAR), organised by the Vrije Universiteit Brussel (Belgium) and the University of Granada (Spain) in 2021. Lozano's research with the *Corpus Escrito del Español L2* (CEDEL2; Lozano 2016, 2021b),[1] for example, focused on how learners of different first languages (L1) produce anaphora in Spanish. However, although these studies consider different language proficiencies, they do not have a pedagogical approach.

Our study contributes to the field by presenting and analysing two new learner corpora: the *Brazilian Learners of Anaphora in English* (BRANEN) and the *Aprendices Brasileños de Anáfora en Español* (BRANES), built to investigate the learning of anaphora under two instructional modes (synchronous and asynchronous). The novelty of this research is to investigate anaphora resolution in a combination of languages that has not been the focus of most studies so far (Brazilian Portuguese L1 and English or Spanish L2), and to collect written data at four points in time (pre-test, post-test 1, post-test 2, delayed post-test) to investigate the learning of nominal, pronominal, and zero anaphora during an online course focused on this mechanism.

A total of 45 participants, who were Brazilian undergraduate students, had an intermediate-to-advanced proficiency level in the foreign language (30 students of English and 15 of Spanish) and were randomly assigned into three groups: one had synchronous lessons on the topic, one had asynchronous lessons, and a third one was the control group (which had no lesson). They all completed short narratives in four moments,

[1] http://cedel2.learnercorpora.com/search

and their written texts were compiled to investigate how a different instructional mode can better contribute to the learning of this specific discourse mechanism.

By the analysis of these corpora, this paper intends to answer the following research questions: (RQ1) Are the new corpora representative of Brazilian learners of English or Spanish? (RQ2) What are the differences between Brazilian learners of English or Spanish on the production of anaphora? (RQ3) What are the differences between the instructional modes (synchronous, asynchronous, and control) on the learning of anaphora?

The outline of the paper is as follows. In Section 2 we discuss the cross-linguistic influence on anaphora resolution, some of the learner corpora available to this purpose, and the effects of instructional modes on the L2 learning of anaphora. In Section 3 we describe the research method of this study and in Section 4 we present our findings and the analysis of the results. The paper closes with some considerations for future investigations.

2. Review of literature

2.1. Cross-linguistic influence on anaphora resolution

As previously explained, anaphora can be defined as a discourse mechanism in which an element in the text (anaphor) refers back to another element (antecedent). There are many types of anaphoric elements, such as nominal, pronominal, and zero anaphora, and their use differs across languages. Pronominal anaphora, for example, is predominant in English, a non-null-subject language, while zero anaphora prevails in Portuguese and Spanish, known as null-subject languages (Chomsky 1981; Rizzi 1982). Still, zero anaphora is commonly used in English in coordinate clauses with the same subject (Quesada and Lozano 2020), something which English learners might overlook when writing in the foreign language.

The anaphoric setup in the learners' L1 can influence their anaphoric behaviour in the L2 and, to avoid ambiguities and misunderstandings, L2 intermediate learners tend to be more explicit than native speakers in their discourse (Hendriks 2003). Although the amount of over-explicitation varies according to the target language, the preference to be redundant rather than ambiguous is related to pragmatic felicity (Lozano 2016). Considering Grice's (1975) second maxim of Quantity (do not make your contribution

more informative than is required), speakers should use null forms whenever possible, as in (2); and, according to the second maxim of Manner (avoid ambiguity), they should prefer redundancy over ambiguity, as in (3).

(2) **Mary** arrived home and ~~**she**~~/**Ø** called Anna.

(3) Mary called **Anna** when ~~**she**~~/**Anna** was travelling.

Focusing on the learning of Spanish by English native speakers, Lozano (2016) analysed the *Corpus Escrito del Español L2* (CEDEL2) and proposed the *Pragmatic Principles Violation Hypothesis*. According to it, very advanced Spanish learners prefer to be redundant by using pronouns to continue a topic, as in (4), than to be ambiguous by omitting the subject when changing it, as in (5). The author also suggests that L1 and L2 speakers tend to use nominal instead of pronominal anaphora to avoid ambiguity when there are same-gender antecedents.

(4) **María** nos llamó cuando **ella** estaba viajando. (redundant) '**Mary** called us when **she** was travelling'.

(5) María llamó a **Ana** cuando **Ø** estaba viajando. (ambiguous) 'Mary called **Anna** when **Ø** was travelling'.

As Miltsakaki (2002) explains, there are many aspects that influence topic continuity and topic shift, some of which are syntactic. In our previous studies (Bruscato and Baptista 2021d, 2022a, 2022b), we tested different anaphoric strategies used by Portuguese, English, and Spanish learners and native speakers when reading ambiguous sentences. We found out that, while English and Spanish native speakers interpret subject pronouns as continuing the topic, Portuguese native speakers interpret them as topic shifters.

Whereas our previous studies analysed data collected with reading questionnaires, the current paper provides more information on the topic by using corpora to analyse written data. The present investigation aims to answer whether Brazilian learners of English or Spanish are more redundant or ambiguous in their own texts and how they learn to reduce these issues over-time.

2.2. The effects of instructional modes on the L2 learning of anaphora

Several studies have investigated learners' knowledge of anaphora, as shown by Ellis (2008: 608–609), and many others have investigated the impact of the instructional mode

on language learning, as reported by Siemens *et al.* (2015). However, as Li (2014) explains, there is practically no research connecting both topics. Liu (2010), for example, investigated whether the type of feedback (implicit or explicit) used in Computer Assisted Language Learning (CALL) would have any impact on the learning of pronominal anaphora in English as L2. The researcher prepared computer exercises on anaphora and asked 28 Chinese adults with an intermediate level of English proficiency to answer them twice. Half of the group read an explanation after each error, while the rest just received a right or wrong feedback. In the end, there was no difference between the groups, probably because they did not have any lessons, the exercises only took half an hour, and were the same both times.

We only found one previous study that investigated the impact of the instructional mode on the learning of anaphora. Li (2014) compared the learning of zero anaphora in Chinese by English native speakers who had onsite or online lessons on the topic, but the effect of the instructional mode to learners' text production was not investigated. Still, the results showed that those who had asynchronous lessons performed better than the others. Possibly an asynchronous course could also improve students' writing, something the current paper intends to answer.

2.3. Learner corpora available

There are many corpora available to study learners' production, such as the *International Corpus of Learner English* (Granger 2003), the *Multilingual Learner Corpus* (Tagnin 2006), the *Corpus of English as a Foreign Language* (COREFL; Lozano *et al.* 2020), and CEDEL2 (Lozano 2021b), as well as some native corpora built to investigate specific types of anaphora, such as *OntoNotes* (Pradhan *et al*. 2007), *Anaphora Resolution and Underspecification* (Poesio and Artstein 2008), and *WikiCoref* (Ghaddar and Langlais 2016). In Portuguese, for example, there are corpora focused on zero (Baptista *et al.* 2016), pronominal (Marques 2013), or nominal anaphora (Pardo *et al.* 2017). However, these are native corpora, and no learner corpus seemed to have the instructional mode as a variable before BRANEN and BRANES.

An example of a multilingual learner corpus that compiled written synchronous and asynchronous computer-mediated communication texts is the *Multilingual Learner Corpus* (MiLC; Andreu *et al*. 2010). However, it was only used to investigate

interlanguage errors in teleconferences and emails, and does not take into account the learning progress in online instructional modes. BRANEN and BRANES, therefore, bring a new perspective to corpora studies by considering different instructional modes in their compilation and analysis.

3. METHOD

3.1. Participants

In this paper we present two corpora, BRANEN and BRANES, which contain texts written by 45 Brazilian undergraduate students with a major in English or Spanish who were in the third or fifth semester of their courses. Texts were collected during a short online course on anaphora in the first semester of 2020.

There were 15 Spanish learners and 30 English learners. Most of them (62%) were in the third semester of their courses, 73 per cent were female, and the average age was 20 (they were between 18 and 41 years). For each language, participants were randomly divided into three groups: one had two synchronous lessons on anaphora, another had two asynchronous lessons, and the control group did not take any lessons.

All participants agreed to take part in the research and answered a grammar questionnaire to ensure they had an intermediate-to-advanced proficiency level in the foreign language. The proficiency test had 20 reading questions, taken from Cambridge University or the Cervantes Institute, equally distributed between levels A2 and C1. Although the English learners' scores were a bit under the Spanish learners' scores, the results among groups were very similar, as can be seen in Table 1 below.

Language	Group	Mean	Standard Deviation
Spanish	Synchronous (N=5)	15	2
	Asynchronous (N=5)	15	2
	Control (N=5)	14	5
English	Synchronous (N=10)	12	4
	Asynchronous (N=10)	14	3
	Control (N=10)	13	5

Table 1: Grammar test results (retrieved from Bruscato and Baptista 2021c)

3.2. Experiment

The university e-learning platform (Moodle) was used during the course. The synchronous groups participated in two videoconference lectures of 90 minutes each, while the asynchronous groups watched short videos, read texts, participated in discussion forums, and answered automatic exercises. As explained in Bruscato and Baptista (2021c: 7), each experimental lesson included:

> activation of prior knowledge on the topic; lecture on anaphora for half an hour; reading and analysis of material; group discussion; reading and writing exercises; and feedback. […] In the first lesson, students introduced themselves; learned about cohesion; the types of anaphora; and the subject, object, and possessive pronouns in the language of study; worked with corpus; completed sentences with the correct pronouns; and did an exercise similar to the test. In the second lesson, they were challenged to solve the ambiguity of some sentences; learned about ambiguity resolution, demonstrative, and relative pronouns; corrected and completed some sentences with pronouns; analysed the coreferences in a fable, comparing their manual analysis with an automatic one; and, again, they did an exercise like the test.

As shown elsewhere (Bruscato and Baptista 2021a, 2022b, 2022c), students wrote short narratives of 100–150 words in four different moments: before the course started (to check students' performance before the intervention), after the first lesson (to measure their progression during the course), after the second lesson (to check their progression when they completed the course), and a month after the course ended (to investigate if students still remembered what they studied). Texts were different but followed the same structure, they were all third-person narrative fictional texts with multiple antecedents.

After reading the beginning of a story (with ten hidden anaphoric problems to solve, previously analysed in Bruscato and Baptista 2021c), students corrected the mistakes they found, and were then asked to conclude the text and submit their files via Moodle. The task was planned to ensure that every student would write about the same topic and that there were multiple antecedents in the texts. The pre-test is presented below.

Instruction: Read the beginning of a narrative and correct the mistakes you find, then write an end to the story between 100 and 150 words.

John and Mary were twins and they were only twelve years old when became orphans. Before these misfortune, John and Mary lived with them parents, Joseph and Ana, that loved they very much. They were all happy, until the country declared war. Joseph was sent to fight, and his wife had to take care of the children and the house. One day, a letter from the government arrived. Ana already knew her content: hers husband was dead. The widow became herself deeply depressed and could not get out of bed. In despair, John and Mary decided to visit the only neighbour they had (they called her witch) to ask for help.

3.3. Corpora

The corpora were first made available on *Sketch Engine*[2] (Kilgarriff *et al*. 2014), a corpus managing and text analysis software, and include metadata about the participants' group (asynchronous, synchronous, control) and testing moment (1, 2, 3, 4). The *Sketch Engine* corpus query system was chosen because it is commonly used by linguists, and because the *European Lexicographic Infrastructure*[3] project provides all academic institutions in the EU free access to the software, at least until 2022.

After their compilation, the corpora were manually annotated using the *Recogito* annotation tool,[4] an online free software that allows the establishment of unilateral, oriented relations between anaphors and antecedents. An anaphora expert annotated the whole corpora, while another expert was responsible for annotating 20 per cent of the texts, which were randomly selected. After the annotation was completed, the intercoder reliability coefficients were calculated using *ReCal2: Reliability for 2 Coders*.[5] The codes used for the anaphors and to establish the anaphoric relation were very similar (agreement around 95%), and the chosen antecedents were the same in about 85 per cent of the time.

Since the first text was written before the course started and we aimed to analyse how learners processed anaphora in comparison to native speakers, three Spanish and six English native speakers also volunteered to do the first task. Their texts were annotated and were made available with the learner corpora.

Based on Lozano's (2016) annotation scheme of subject expressions, third-person human subjects of finite clauses and their antecedents were annotated following the

[2] www.sketchengine.eu
[3] https://elex.is/
[4] https://recogito.pelagios.org/
[5] http://dfreelon.org/utils/recalfront/recal2/#doc

scheme shown in Table 2. First, the form of the expression was annotated. Since blank spaces cannot be marked on *Recogito*, in case of zero anaphora the annotation was on the primary verb. All anaphors were subjects, but antecedents could also be non-subjects. Then, they received a tag according to the type of clause they were in, and there was an option to annotate if the expressions were ambiguous or redundant. After that, the intrasentential and intersentential relations were established. When necessary, there was the possibility to specify if it was a case of cataphora or a partial relation.

The annotation scheme with the tags and examples from the corpora are presented in Table 2.[6]

Form	
Zero – *zr*	She was good, generous, and helped other people [...]. ia1e[1]
Pronoun – *pp*	[...] they would get what they wanted [...]. Ia1c
Determiner – *dt*	[...] the two started [...]. Ia3c
Common noun – *nc*	[...] the twins arrived home late [...]. Ia1c
Proper noun – *np*	Mary knocked on the door [...]. Ia1a
Function	
Subject – *sj*	They missed their dad [...]. ia1c
Non-subject – *ns*	The witch told them that they had to [...]. ia1c
Clause	
Main clause – *mc*	They went there for help [...]. ia1c
Coordinate clause – *cc*	They missed their dad and they were worried about their mom [...]. ia1c
Subordinate clause – *sc*	The witch told them that they had to [...]. ia1c
Pragmatic-felicity (optional)	
Ambiguous – *am*	She also tried to help the kids with their mother, but she ended up very sick [...]. is1d
Redundant – *rd*	They missed their dad and they were worried about their mom [...]. ia1c
Relation	
Intrasentential – *in*	They missed their dad and they were worried about their mom [...]. ia1c
Intersentential – *tr*	They went there for help. They wanted their old life back [...]. ia1c
Cataphora (optional) – *ca*	Since she was a kid, Mary knew [...]. ic1j
Partial (optional) – *pr*	She seemed happy to help their mother, so they all went to the bedroom where Anna was [...]. ia1b

Table 2: Annotation scheme with tags and examples

Possessive, reflexive, and relative pronouns were not annotated in this phase. This study focuses on subject expressions and their antecedents and, since relative pronouns appear right after their antecedent in the text, they were not relevant for the current purposes.

Figures 1 and 2 show examples of annotated texts. As can be seen, the English text had many subject pronouns (e.g. ***They*** *went there for help.* ***They*** *wanted their old life*

[6] The code after each example indicates a specific file in the corpora.

back. ***They*** *missed their dad and* ***they*** *were worried about their mom.*). The Spanish excerpt, on the contrary, had no subject pronoun, but exhibited many cases of zero and nominal anaphora (e.g. *La* ***bruja*** *había puesto veneno en la sopa y se* ***reía***. 'The **witch** had added poison to the soup and **was laughing**.'). After the annotation was completed, results were exported into .csv files and analysed in SPSS (v. 26; IBM Corporation 2020).

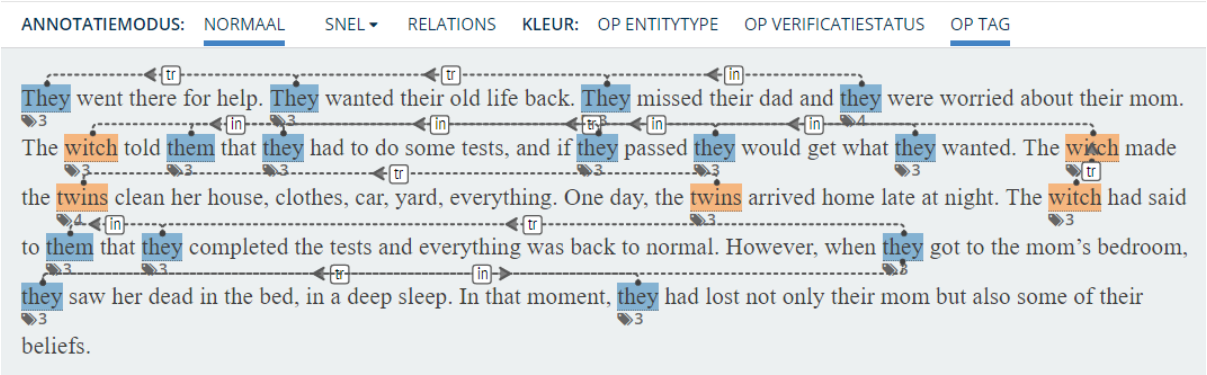

Figure 1: Subject anaphora annotation in English narrative

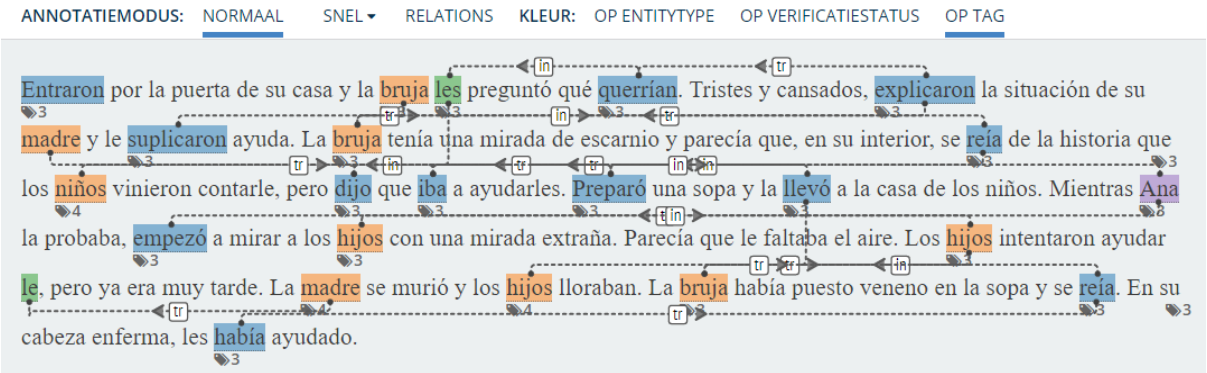

Figure 2: Subject anaphora annotation in Spanish narrative

BRANEN has 120 documents and was automatically part-of-speech (POS) tagged by *Sketch Engine* with the *Modified English TreeTagger*, while BRANES has 60 documents and was POS-tagged with the *Spanish FreeLing* tagset. Table 3 presents the size of the corpora. More information about them can be found in Bruscato and Baptista (2021c).

	documents	**sentences**	**lemmas**	**unique words**	**words**	**tokens**
BRANEN	120	1,069	1,678	2.242	17,454	19,934
BRANES	60	543	1,299	2.095	9,021	10,233

Table 3: Size of the corpora (data retrieved from *Sketch Engine*)

4. RESULTS AND DISCUSSION

The results retrieved from the corpus-based analysis will be presented and discussed in what follows. Before that, however, some descriptive information about the anaphoric relations in the corpora and the distribution of the anaphoric forms will be provided. Tables 4 and 5 below present the number of anaphoric relations per group and test moment.

	BRANEN	**BRANES**
Asynchronous	616 (35.9%)	310 (33.6%)
Control	516 (30.1%)	285 (30.9%)
Synchronous	582 (34%)	327 (35.5%)
Total	1,714 (100%)	922 (100%)

Table 4: Number of anaphoric relations per group

	BRANEN	**BRANES**
Text 1	427 (24.9%)	234 (25.4%)
Text 2	397 (23.1%)	202 (21.9%)
Text 3	418 (24.4%)	235 (25.5%)
Text 4	472 (27.5%)	251 (27.2%)
Total	1,714 (100%)	922 (100%)

Table 5: Number of anaphoric relations per test moment

The differences in the number of anaphoric relations among groups (Table 4) was negligible in both corpora, since there was only around a five per cent difference between the minimum and maximum values. The same was found when considering the test moments (Table 5). Still, the control groups established fewer relations than the others in both corpora. Considering the four test moments, which were done under similar conditions, there were slightly fewer anaphoric relations in the second test and an above-average number of relations in the fourth, but the number of anaphoric relations was highly correlated in the two corpora among groups ($r = 0.918$) and test moments ($r = 0.953$). This was already expected, since all participants were instructed to write a similar number of words (100–150) in each text and were provided with the same number of possible antecedents.

After identifying a similar number of anaphoric relations among groups and test moments, we compared the distribution of the anaphoric forms used by English and Spanish learners and by native speakers in the first test (Table 6).

	BRANEN		BRANES	
	Learners (n=30)	Natives (n=6)	Learners (n=15)	Natives (n=3)
Zero	42 (9.8%)	14 (12.7%)	128 (54.7%)	22 (56.4%)
Determiner	4 (0.9%)	1 (0.9%)	8 (3.4%)	0 (0%)
Pronoun	233 (54.6%)	56 (50.9%)	25 (10.7%)	1 (2.6%)
Common noun	97 (22.7%)	21 (19.1%)	43 (18.4%)	11 (28.2%)
Proper noun	51 (11.9%)	18 (16.4%)	30 (12.8%)	5 (12.8%)
Total	427 (100%)	110 (100%)	234 (100%)	39 (100%)

Table 6: Distribution of the anaphoric forms in the first test

The distribution of the anaphoric forms in the first test between learners and native speakers from BRANEN and BRANES was very similar (r(EN) = 0.988; r(ES) = 0.962). As Table 6 shows, determiners were hardly used as anaphors (0 to 3.4%) and, in each language, there was a preferred type of anaphora. As expected, in English, more than half of the subject anaphors were pronouns (50.9%), while in Spanish ellipsis was preponderant (56.4%). However, L2 learners produced slightly more pronominal anaphora (+3.7% in English and +8.1% in Spanish) than native speakers. In general, Spanish learners also used less nominal anaphora (-9.8%), while English learners used more common nouns as subjects than natives did (+3.6%).

The similarity between learners and native speakers can be explained by the students' intermediate-to-advanced level of proficiency in the language. Still, there were some slight differences between the groups, showing that students could sometimes replace pronouns with other types of anaphors.

4.1. Representativeness

Since BRANES had a small number of informants (15 Spanish learners and three native speakers), we compared our results with data from CEDEL2 (Lozano 2021b), namely in the use of zero and pronominal anaphora by L1 European Portuguese (L2 Spanish and L1 Spanish adults). For this, the Chaplin task was used, in which participants had to narrate a silent Charles Chaplin video clip. This corpus consists of 137 written texts from native speakers and 85 from learners, of which 96.5 per cent had an intermediate-to-advanced level in the L2, that is, a proficiency level similar to that of the subjects in our study.

For the use of pronominal anaphora, we first looked for instances of third-person nominative personal pronouns in CEDEL2, but not a single occurrence was found. We then checked if there was any nominative pronoun in the corpus, and there were two

occurrences of first-person personal pronouns from learners. One of the sentences was [...] *él encuentra un billete que dice "cuidame, yo soy huerfano*" [...] ('[...] he finds a note that says "take care of me, I am an orphan" [...]'). Irrespective of the occurrence of pronoun *yo* (*I*), clearly, there was a third-person nominative personal pronoun in this sentence: *él* (*he*). However, its case had not been annotated. Since we could not automatically distinguish nominative from other types of personal pronouns using the tool, we left this search for further research. However, other studies interested in this can download the corpus from the CEDEL2 website and manually annotate it.

To compare the use of zero and pronominal anaphora, we checked the number of occurrences of a punctuation mark or a conjunction followed by a third-person verb with either an ellipsis in the middle, as in (6), or the lemma *él* (*he*), as in (7).

(6) **Sigue** caminando **y pide** a un hombre que lo sujete por un momento [...]. 'He keeps walking and asks a man to hold him for a moment [...]'.

(7) [...] **cuando ella ve** el nene en su cochecito**, ella corre** en dirección a Chaplin [...]. '[...] when she sees the baby in his stroller, she runs to Chaplin [...]'

Table 7 compares the number of zero and pronominal anaphora in BRANES's pre-test (before the intervention) with the frequencies of comparable patterns found in CEDEL2 (for which no intervention took place). Coincidentally, the percentages found of zero and pronominal anaphora compared to their total were identical between the two groups of learners, and extremely similar between the groups of natives. Despite the small number of participants in BRANES, the similar results found in CEDEL2 give some assurance about the remarks made above. In the next subsections, the data in BRANES and BRANEN are detailed and compared in depth.

	BRANES		**CEDEL2**	
	Learners (n=15)	Natives (n=3)	Learners (n=85)	Natives (n=137)
Zero anaphora	128 (83.7%)	22 (96%)	231 (83.7%)	1,187 (97.5%)
Pronominal ana.	25 (16.3%)	1 (4%)	45 (16.3%)	30 (2.5%)
Total	153 (100%)	23 (100%)	276 (100%)	1,217 (100%)

Table 7: Zero and pronominal anaphora in BRANES and CEDEL2

4.2. Differences between Brazilian learners of English or Spanish on the production of anaphora

To answer the second research question, we compare how participants produced anaphora in the first test (pre-intervention). Table 8 presents the frequencies of the pre-test and shows that, in BRANEN and BRANES, both learners and native speakers established anaphoric relations using the different strategies in a similar way (r(EN) = 0.994; r(ES) = 0.957). This result was already expected, due to the students' proficiency in the language.

		BRANEN		**BRANES**	
		Learners (n=30)	Natives (n=6)	Learners (n=15)	Natives (n=3)
Anaphor clause	Main	195 (45.7%)	59 (53.6%)	104 (44.4%)	13 (33.3%)
	Coordinate	115 (26.9%)	27 (24.6%)	63 (27%)	15 (38.5%)
	Subordinate	117 (27.4%)	24 (21.8%)	67 (28.6%)	11 (28.2%)
Anaphor pragmatics	No problem	391 (91.6%)	110 (100%)	176 (75.2%)	34 (87.2%)
	Ambiguous	9 (2.1%)	0 (0%)	6 (2.6%)	0 (0%)
	Redundant	27 (6.3%)	0 (0%)	52 (22.2%)	5 (12.8%)
Anaphoric relation	Intrasentential	176 (41.2%)	40 (36.4%)	107 (45.7%)	19 (48.7%)
	Intersentential	251 (58.8%)	70 (63.6%)	127 (54.3%)	20 (51.3%)
Antecedent function	Subject	313 (73.3%)	86 (78.2%)	161 (68.8%)	31 (79.5%)
	Non subject	114 (26.7%)	24 (21.8%)	73 (31.2%)	8 (20.5%)

Table 8: Frequencies of the first test

Learners behaved almost the same, despite the target language. In general, the differences in their results are less than 5 per cent. Nonetheless, compared to native speakers, some distinctions were found. In learners' texts, almost half of the subject anaphors were in main clauses (46.7% EN; 44.4% ES). In comparison, native English speakers used 7.9 per cent more subject anaphors in main clauses and Spanish speakers used 11.1 per cent fewer subject anaphors in main clauses. In English, there was almost no difference in coordinate clauses, but, in subordinate clauses, natives produced slightly fewer subject anaphors (-5.6%). In Spanish, on the other hand, the results from subordinate clauses were very similar, but, in coordinate clauses, native speakers used 11.5 per cent more subject anaphors than learners did.

Possibly, learners were more influenced by their L1 syntax than by the L2 when using the sentences and therefore behaved the same despite the target language regarding the anaphor's clauses. This is in line with Bruscato and Baptista (2021d, 2022a, 2022b) regarding the anaphoric strategies used by learners when reading. Based on these results,

it was also found that the preferences in English and Spanish as L1 regarding the distribution of anaphoric subjects in the types of clauses seem to differ. While, for example, in English 53.6 per cent of the subject anaphors were produced in main clauses, in Spanish that number decreased to 33.3 per cent. In further research, it would be relevant to compare these results with data from more informants.

Another difference among the English and Spanish texts is the distance between the anaphors and their antecedents. Although learners and native speakers from each language behaved similarly, English native speakers showed a clearer preference to retrieve intersentential antecedents (63.6%, compared to 51.3% in Spanish). This could be related to the previously discussed higher number of anaphoric subjects in English main clauses.

Besides the preference to select intersentential antecedents, most of them were also subjects among native speakers of English. Although the tendency for topic continuity was already expected, Spanish learners chose a subject antecedent 10.7 per cent less frequently than native speakers and, considering all types of anaphora, they were also 9.4 per cent more redundant. 35 out of their 52 occurrences of redundancy were subjects in a main clause, and nine of these were nouns that retrieved the subject from another sentence, as in (8). Considering the other groups (English learners and all natives), there was not much redundancy in general and even less ambiguity.

(8) Los **niños**, que también se apegaron a la vecina, muy agradecidos, aceptaron la propuesta. Y aunque tristes, los **niños** estaban muy agradecidos por la compasión y la empatía de su vecina [...] (ec1b). 'The **children**, who also attached themselves to the neighbour, very grateful, accepted the proposal. And although sad, the **children** were very grateful for the compassion and empathy of their neighbour [...]'

Since a substantial part of the anaphors found in the corpus recover non-subject antecedents, these values call for further analysis. This is the main purpose of Tables 9, 10, and 11, which present the results per group for main, coordinate, and subordinate clauses, respectively.

Table 9 shows the results for anaphora in main clauses. As expected, in English, either nominal or pronominal subjects are used in main clauses. For both groups of informants, there were around 33 per cent of nouns and 39 per cent of pronouns that recovered a subject antecedent. Since English is not a null-subject language, zero

anaphora in subject position of main clauses is not grammatical, and learners complied with this general rule.

Anaphor form	Antecedent function	BRANEN		BRANES	
		Learners (n=30)	Natives (n=6)	Learners (n=15)	Natives (n=3)
Zero	Subject	0 (0%)	0 (0%)	34 (32.7%)	5 (38.5%)
	Non-subject	0 (0%)	0 (0%)	2 (1.9%)	0 (0%)
Determiner	Subject	2 (1%)	0 (0%)	2 (1.9%)	0 (0%)
	Non-subject	1 (0.5%)	0 (0%)	2 (1.9%)	0 (0%)
Pronoun	Subject	76 (39%)	23 (39%)	9 (8.7%)	0 (0%)
	Non-subject	19 (9.7%)	7 (11.9%)	4 (3.9%)	0 (0%)
Noun	Subject	62 (31.8%)	21 (35.6%)	23 (22.1%)	6 (46.1%)
	Non-subject	35 (18%)	8 (13.5%)	28 (26.9%)	2 (15.4%)
Total		195 (100%)	59 (100%)	104 (100%)	13 (100%)

Table 9: Anaphora in main clauses

In Spanish, either nominal or zero anaphora are used. While practically all ellipses recovered a previous subject, nouns were used to recover both, subject and non-subject antecedents. Around a fifth of learners' subject anaphors were nouns that retrieved a previous subject and, as mentioned before, nine of these were redundant. Unlike native speakers, Spanish learners produced pronominal anaphora (but to a lesser extent than English learners).

Table 10 presents the results for coordinate clauses. In coordinate clauses, both languages retrieve intrasentential subjects by zero anaphora. However, English native speakers used 15.4 per cent more ellipses than learners did, while Spanish learners used it 16.2 per cent more frequently than native speakers. Pronominal anaphora was also common in English for the same task, especially for learners who, as already mentioned, used fewer ellipses. To select intersentential antecedents, all groups mostly used nominal anaphora, as in example (8), above.

Anaphor form	Anaphoric relation	Antecedent function	BRANEN		BRANES	
			Learners (n=30)	Natives (n=6)	Learners (n=15)	Natives (n=3)
Zero	Intra.	Subject	42 (36.5%)	14 (51.9%)	48 (76.2%)	9 (60%)
		Non-subject	0 (0%)	0 (0%)	2 (3.2%)	2 (13.3%)
	Inter.	Subject	0 (0%)	0 (0%)	2 (3.2%)	0 (0%)
Determiner	Intra.	Non-subject	0 (0%)	1 (3.7%)	2 (3.2%)	0 (0%)
Pronoun	Intra.	Subject	27 (23.5%)	5 (18.5%)	1 (1.5%)	0 (0%)
		Non-subject	16 (13.9%)	2 (7.4%)	2 (3.2%)	0 (0%)
	Inter.	Subject	4 (3.5%)	1 (3.7%)	0 (0%)	0 (0%)
		Non-subject	1 (0.8%)	0 (0%)	0 (0%)	0 (0%)
Noun	Intra.	Subject	4 (3.5%)	0 (0%)	0 (0%)	0 (0%)
		Non-subject	4 (3.5%)	1 (3.7%)	0 (0%)	0 (0%)
	Inter.	Subject	13 (11.3%)	3 (11.1%)	4 (6.3%)	3 (20%)
		Non-subject	4 (3.5%)	0 (0%)	2 (3.2%)	1 (6.7%)
	Total		115 (100%)	27 (100%)	63 (100%)	15 (100%)

Table 10: Anaphora in coordinate clauses

Finally, the results for subordinate clauses are shown in Table 11. In subordinate clauses, there is also a preference to retrieve a subject antecedent, which is usually intrasentential. While Spanish speakers prefer to use zero anaphora, English speakers mainly use pronominal anaphora for that matter. However, there was some difference between the English groups. Natives chose pronouns to recover 11.5 per cent more intrasentential subjects than learners, who chose them to select 9.6 per cent more intrasentential non-subjects. It seems there is a stronger preference for topic continuity in English as L1. Lastly, as already seen before, nominal anaphora tends to select intersentential antecedents.

Anaphor form	Anaphoric relation	Antecedent function	BRANEN		BRANES	
			Learners (n=30)	Natives (n=6)	Learners (n=15)	Natives (n=3)
Zero	Intra.	Subject	0 (0%)	0 (0%)	22 (32.8%)	3 (27.2%)
		Non-subject	0 (0%)	0 (0%)	10 (14.9%)	1 (9.1%)
	Inter.	Subject	0 (0%)	0 (0%)	7 (10.4%)	2 (18.2%)
		Non-subject	0 (0%)	0 (0%)	1 (1.5%)	0 (0%)
Determiner	Inter.	Subject	1 (0.9%)	0 (0%)	1 (1.5%)	0 (0%)
		Non-subject	0 (0%)	0 (0%)	1 (1.5%)	0 (0%)
Pronoun	Intra.	Subject	45 (38.5%)	12 (50%)	3 (4.5%)	0 (0%)
		Non-subject	21 (17.9%)	2 (8.3%)	6 (9%)	0 (0%)
	Inter.	Subject	19 (16.2%)	3 (12.5%)	0 (0%)	0 (0%)
		Non-subject	5 (4.3%)	1 (4.2%)	0 (0%)	1 (9.1%)
Noun	Intra.	Subject	1 (0.9%)	1 (4.2%)	0 (0%)	2 (18.2%)
		Non-subject	2 (1.7%)	0 (0%)	4 (6%)	0 (0%)
	Inter.	Subject	17 (14.5%)	3 (12.5%)	5 (7.5%)	1 (9.1%)
		Non-subject	6 (5.1%)	2 (8.3%)	7 (10.4%)	1 (9.1%)
	Total		117 (100%)	24 (100%)	67 (100%)	11 (100%)

Table 11: Anaphora in subordinate clauses

As expected, in English zero anaphora was only used in coordinate clauses to select the subject of the previous clause. In Spanish, this was the case for around 40 per cent of ellipses, but, regardless of the clauses, more than 86 per cent of them were used to select a subject antecedent.

Participants mostly used subject pronouns (usually in a main clause) to retrieve subject antecedents, except for Spanish native speakers, who only used one subject pronoun in a subordinate clause to retrieve an intersentential non-subject antecedent, as shown in (9). The majority of anaphoric common and proper nouns were in main clauses, and they also recovered another subject. Spanish learners, however, used nominal anaphora mainly to retrieve non-subject antecedents.

(9) Una vez con el estómago lleno, le contaron la tragedia a la **vecina**, la cual sin dudarlo un momento los invitó a vivir con ella. Su esposo y dos hijos habían muerto en la guerra, así que **ella** también estaba sola. (sn3) ‘Once with a full stomach, they told the **neighbour** about the tragedy, who without doubting for a moment invited them to live with her. Her husband and two children had died in the war, so **she** was alone too.’

In summary, there were many similarities between Brazilian learners of English and Spanish learners of English in the production of anaphora. To start, around 45 per cent of the anaphors were in main clauses, which also meant that more than 50 per cent of the anaphors had an intersentential antecedent.

As expected, most anaphors in both languages (around 70%) continued the topic by retrieving the previous subject. More than half of the subject anaphors were pronouns in English and ellipsis in Spanish, and more than 70 per cent of them were used for topic continuity. Nouns and pronouns were used to shift the topic, and most cases (more than three fourths) did not have any pragmatic issues (i.e. ambiguity and redundancy). However, more than one fifth of the anaphors produced by Spanish learners were considered redundant. We noted that, in spite of the preferences for pronominal or zero anaphora, English and Spanish learners behaved in a similar way. Still, there were some differences between them and native speakers. English native speakers used more main clauses, while Spanish native speakers preferred to coordinate clauses. Compared to native speakers, it was also clear that learners could use more frequently zero anaphora instead of pronominal anaphora in English coordinate clauses and in Spanish main clauses to continue the topic. This was addressed during the online course and will be discussed in Section 4.3.

4.3. Differences between the instructional modes (synchronous, asynchronous, and control) on the learning of anaphora

To answer the third research question, we will discuss some differences in the use of anaphora by the experimental and control groups over time. We will first analyse the English groups and then the Spanish ones.

4.3.1. The English groups

As stated in previous sections, English learners were not ambiguous and were not much redundant in their texts. Still, they were instructed about how to omit the subject expression when possible and, as Figure 3 shows, the number of redundant anaphors decreased for both experimental groups.

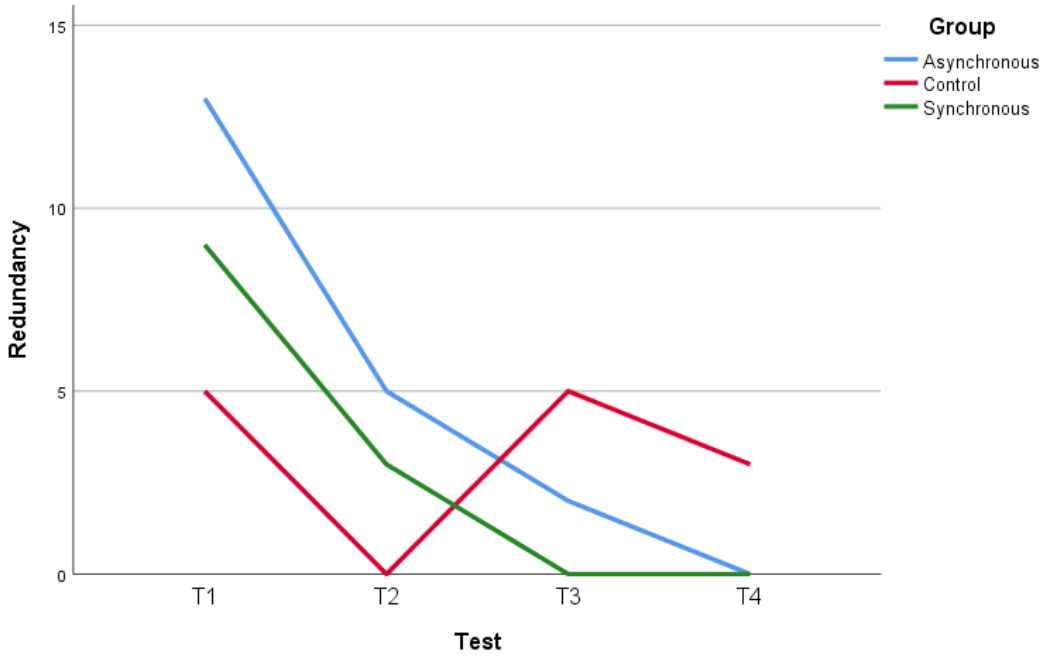

Figure 3: Redundancy in English

Although English learners were not so redundant in the first test compared to native speakers, they could still learn to use more zero anaphora in coordinate clauses with the same subject. They studied how to do it during the course and changed their anaphoric behaviour.

Figures 4 and 5 show that the asynchronous group started to use more zero anaphora in coordinate clauses, as well as less pronominal anaphora in main clauses to continue the topic, especially until the second post-test. This possibly happened because this group of learners chose to coordinate more clauses instead of separating them in different sentences.

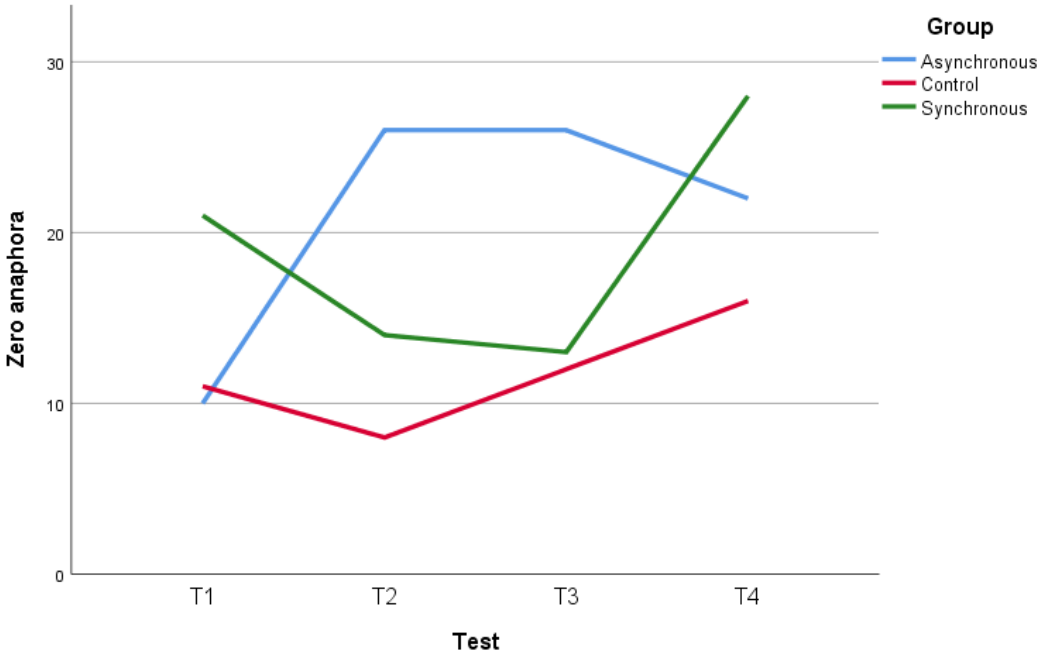

Figure 4: Zero anaphora in English coordinate clauses to retrieve the previous subject

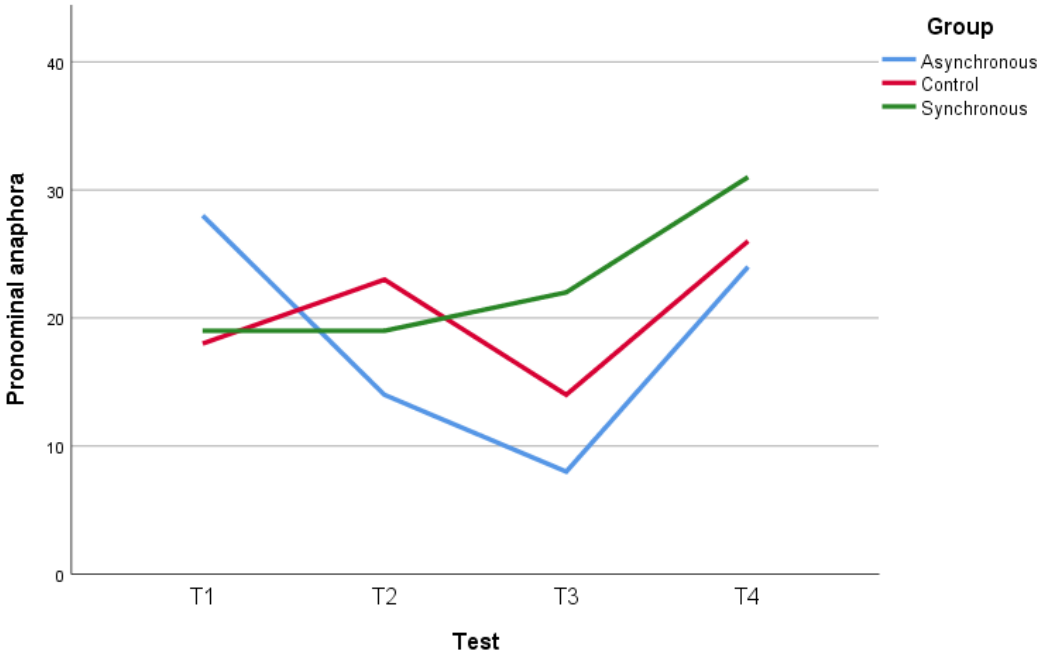

Figure 5: Pronominal anaphora in English main clauses to retrieve a previous subject

The data in Figure 6 also indicate a decrease in the use of pronominal anaphora by the synchronous group in coordinate clauses when there was topic continuity, but, as with the asynchronous group, the changes happened mainly until the third test. In the final test, one month after the course, the results were more similar to the pre-test.

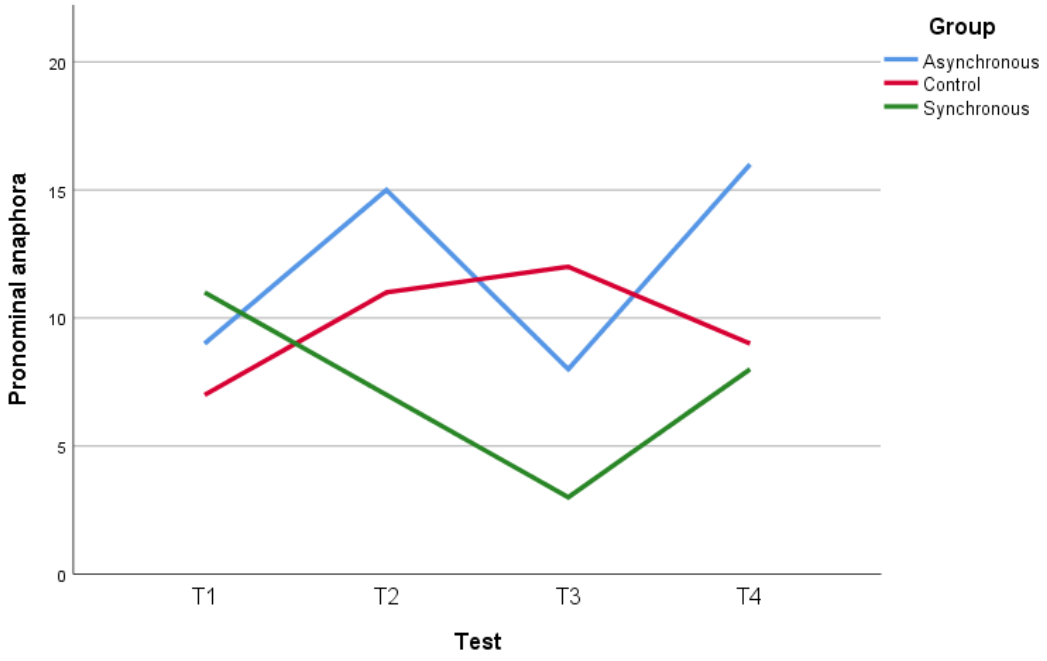

Figure 6: Pronominal anaphora in English coordinate clauses to retrieve the previous subject

In view of these findings, we can conclude that both experimental groups learned to be less redundant and to use less pronominal anaphora for topic continuity. However, only the asynchronous groups started to use more zero anaphora, and the changes were mainly until the second post-test.

4.3.2. The Spanish groups

In general, Spanish learners behaved very similarly to native speakers, probably because Portuguese and Spanish are both null-subject languages. As proposed by Lozano (2016) and confirmed in our study, learners were not ambiguous, but redundant in their texts. To solve this issue, during the course they studied how to use more zero anaphora instead of pronominal anaphora when continuing the topic and thus reduced the number of redundant anaphors, as can be seen in Figure 7.

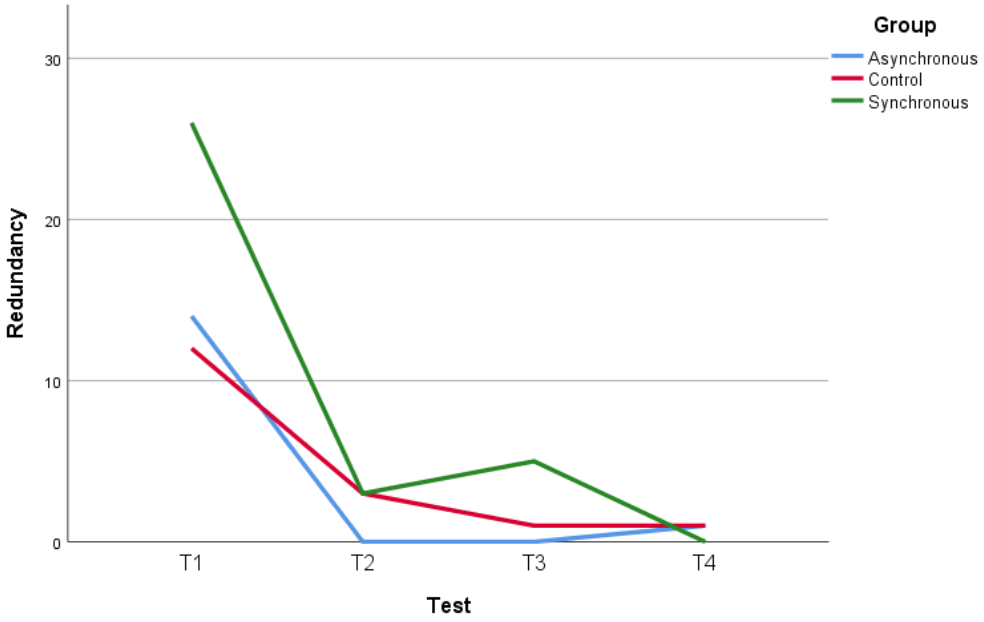

Figure 7: Redundancy in Spanish

Based especially on the results from main clauses, Spanish learners could also learn to use more zero instead of pronominal anaphora. They studied it during the course and, as Figures 9 and 9 present, the asynchronous group increased the use of zero anaphora until test 3, as well as continuously decreased the use of pronominal anaphora. Although the synchronous group also used fewer pronouns in test 2, the numbers increased in the following tests.

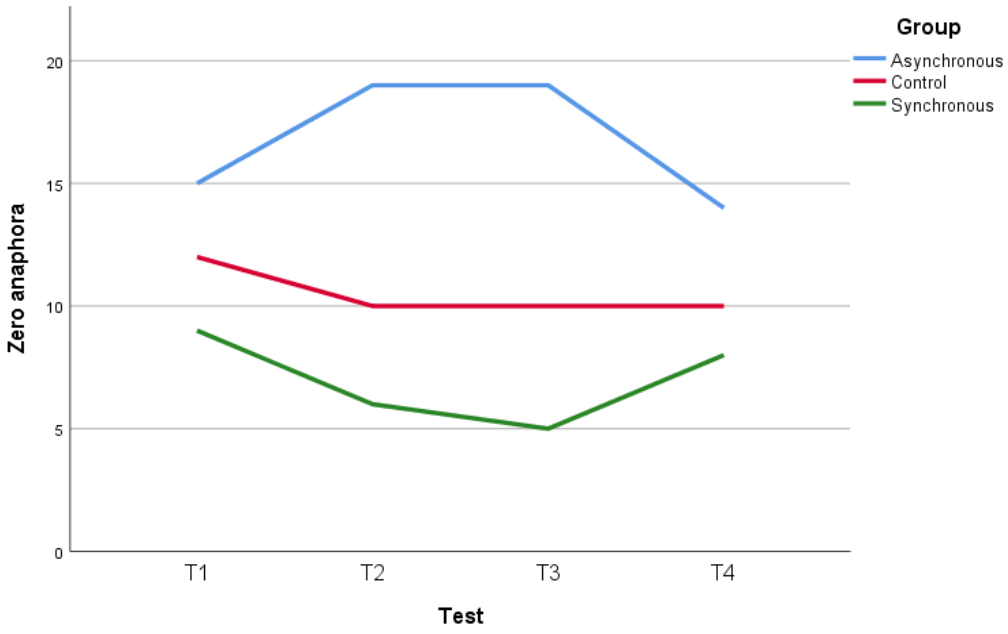

Figure 8: Zero anaphora in Spanish main clauses

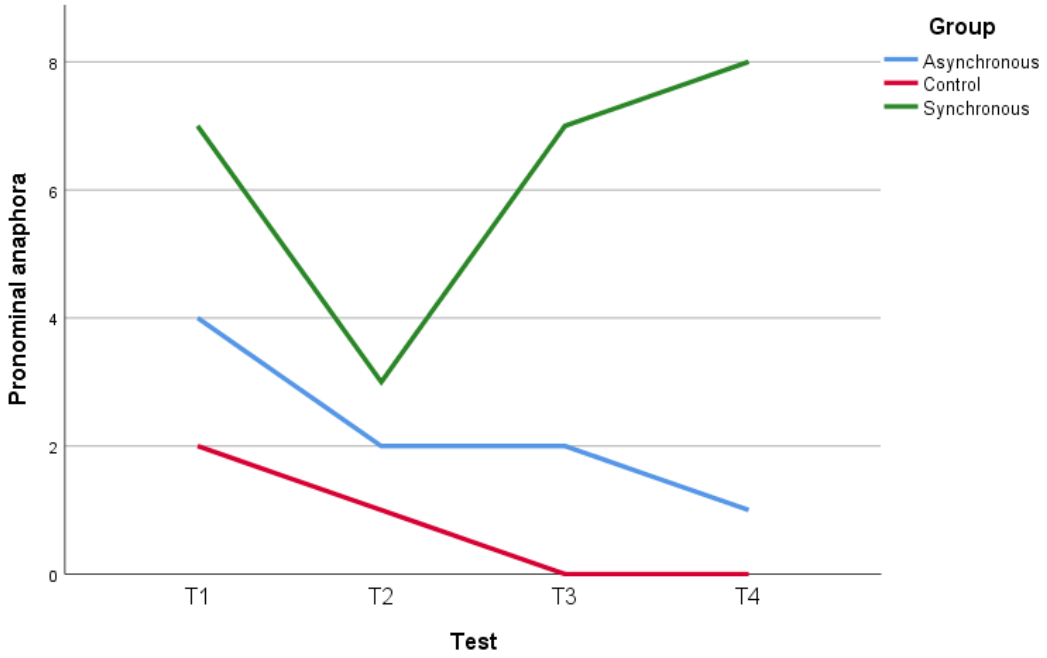

Figure 9: Pronominal anaphora in Spanish main clauses

In general, the asynchronous group performed better than the other groups. Regarding the increase of zero anaphora in Spanish main clauses and English coordinate clauses, the asynchronous group differed from the synchronous and control groups in the second and third tests.[7] The synchronous and control groups did not present a significant difference between each other, and the three groups behaved similarly in the initial pre-test and the final post-test.

5. Conclusion

The present paper investigated the use of nominal, pronominal, and zero anaphora in two written corpora: BRANEN and BRANES. We designed an online course in two different instructional modes (synchronous and asynchronous) to investigate their impact on the learning of anaphora in English and Spanish over time. There were 45 participants (including control groups) who wrote narrative texts in four moments. Based on Lozano's (2016) annotation scheme of subject expressions, we annotated manually third-person

[7] The Kruskal-Wallis test performed in SPSS (v. 26) only identified statistically relevant differences from the asynchronous group and the other groups on the second [$X^2(2) = 6.234$; $p = 0.044$] and third tests [$X^2(2) = 8.054$; $p = 0.018$]. The limited sample size does not allow for further elaboration.

human subjects of finite clauses and their antecedents in the texts using the *Recogito* annotation platform.

Since there was a small number of Spanish speakers (15 learners and three natives), we compared the use of zero and pronominal anaphora in BRANES and CEDEL2 (Lozano 2021b) before we started analysing and interpreting the results. Coincidentally, the percentages found were identical between the two groups of learners, and extremely similar between the groups of natives. Thus, we could answer RQ1 and consider our corpora representative.

After attesting the representativeness of the corpora, we analysed how Brazilian speakers processed anaphora in English and Spanish as foreign languages before the intervention (Tables 6 to 11) to answer the RQ2. We found similarities between learners and native speakers, which could be explained by the apprentices' intermediate-to-advanced level of proficiency in the language, but learners' distribution of anaphora in the types of clauses was much alike, regardless of the target language. It is possible that they have been more influenced by their L1 syntax than by the L2 when writing the sentences (as already suggested by Bruscato and Baptista 2021d, 2022a, 2022b regarding learner's reading strategies). Our study also confirmed Lozano's (2016) hypothesis, according to which learners are more redundant than ambiguous. In the pre-test, participants could have used more zero anaphora instead of pronominal anaphora in English coordinate clauses and in Spanish main clauses to continue the topic.

Finally, we investigated the effect of the instructional modes (synchronous and asynchronous) on the learning of the discursive mechanism (Figures 3 to 9) to answer RQ3. Although both experimental groups showed progress on the learning of anaphora, contrary to the control groups, the results revealed that the asynchronous instructional mode was more effective, probably because learners had more opportunities to read and write on the written forums than the synchronous groups on the oral discussions, but only until the third test.

In spite of the interesting remarks made above, the current study had some limitations that must be acknowledged. Firstly, the corpora contain only 180 short texts written by 45 learners with an intermediate-to-advanced level in the foreign language, a somewhat limited sample considering all possible target subjects. It would be interesting to compare the results here with data from more informants and with different levels of

proficiency. Besides, this was a short course, and the experimental groups had only two lessons on anaphora. In the future, the duration of the course could be extended, and it could include more testing moments. The effectiveness of the course over a longer period could also be investigated. Finally, our research focused only on third-person human subject anaphors, but non-human subjects or even verb complements could be annotated and analysed. To this end, the data retrieved from BRANEN and BRANES can be put to good use.

The major contribution of this paper is to show that Brazilian learners of Spanish and English use anaphora differently in relation to their instructional mode and to provide the scientific community with real, textual data for further investigation. To the best of our knowledge, a distant learning mode-specific approach to anaphora learning like that had not been described yet. In the future, besides pursuing some of the lines of research already sketched above, we also plan to investigate the impact of synchronous and asynchronous learning to the understanding and the production of anaphora in spoken texts.

REFERENCES

Andreu-Andrés, Mª Ángeles, Aurora Astor, María Boquera, Penny MacDonald, Begoña Montero and Carmen Pérez. 2010. Analysing EFL learner output in the MiLC project: An error it's*, but which tag. In Mari Campoy-Cubillo, Begoña Bellés-Fortuño and María Lluisa Gea-Valor eds. *Corpus-based Approaches to English Language Teaching*. London: Continuum, 167–179.

Baptista, Jorge, Simone Pereira and Nuno Mamede. 2016. ZAC: Zero Anaphora Corpus. In João Silva, Ricardo Ribeiro, Paulo Quaresma, André Adami and António Branco eds. *Proceedings of the International Conference on Computational Processing of the Portuguese Language 2016*. Tomar: Springer International Publishing, 38–45.

Bruscato, Amanda Maraschin and Jorge Baptista. 2021a. BRANEN and BRANES Corpora. In *Proceedings of the 9th European Conference on Language Learning*, London: IAFOR, 1–13. https://doi.org/10.22492/issn.2188-112X.2021.3

Bruscato, Amanda Maraschin and Jorge Baptista. 2021b. Designing an online course to teach anaphora in foreign languages. *International Journal of Second and Foreign Language Education* 1/1: 1–9.

Bruscato, Amanda Maraschin and Jorge Baptista. 2021c. Synchronous and asynchronous distance learning of anaphora in foreign languages. *Texto Livre* 14/1: 1–16.

Bruscato, Amanda Maraschin and Jorge Baptista. 2021d. Resolução de ambiguidade anafórica em português, inglês e espanhol (estudo-piloto). *Alfa* 65/1: 1–22.

Bruscato, Amanda Maraschin and Jorge Baptista. 2022a. Resolução de ambiguidade anafórica em português, inglês e espanhol: Comparação de dados obtidos com falantes de PE e PB. *Domínios de Lingu@gem* 16/1: 1–29.

Bruscato, Amanda Maraschin and Jorge Baptista. 2022b (in press). The resolution of

ambiguous anaphora in English, Spanish, and Portuguese. *Delta*. https://doi.org/10.1590/1678-460x202151581

Chomsky, Noam. 1981. *Lectures on Government and Binding: The Pisa Lectures*. Dordrecht: Foris.

Ellis, Rod. 2008. *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Ghaddar, Abbas and Philippe Langlais. 2016. Wikicoref: An English coreference-annotated corpus of Wikipedia articles. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the Language Resources and Evaluation Conference*, LREC 2016. Portorož: European Language Resources Association, 136–142.

Granger, Sylviane. 2003. The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly* 37/3: 538–546.

Grice, Herbert Paul. 1975. Logic and conversation. In Peter Cole and Jerry Morgan eds. *Syntax and Semantics, Vol. 3: Speech Acts*. New York: Academic Press, 41–58.

Hendriks, Henriëtte. 2003. Using nouns for reference maintenance: A seeming contradiction in L2 discourse. In Anna Ramat ed. *Typology and Second Language Acquisition*. Berlin: Mouton de Gruyter, 291–326.

IBM Corporation. 2020. *IBM SPSS Statistics for Windows, version 26*. Armonk: IBM Corporation.

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

Li, Liu. 2014. Computer-assisted vs. classroom instruction on developing reference tracking skills in L2 Chinese. In Shuai Li and Peter Swanson eds. *Engaging Language Learners through Technology Integration: Theory, Applications, and Outcomes*. Hershey: Information Science Reference, 72–96.

Liu, Rong. 2010. *The Acquisition and Online Processing of Anaphora by Chinese-English Bilinguals: A Computer Assisted Study*. Arizona: University of Arizona dissertation.

Lozano, Cristóbal. 2016. Pragmatic principles in anaphora resolution at the syntax-discourse interface: Advanced English learners of Spanish in the CEDEL2 corpus. In Margarita Alonso-Ramos ed. *Spanish Learner Corpus Research: Current Trends and Future Perspectives*. Amsterdam: John Benjamins, 235–265.

Lozano, Cristóbal. 2021a. Anaphora resolution in Second Language Acquisition. In Mark Aronoff ed. *Oxford Bibliographies in Linguistics*. Oxford: Oxford University Press. https://www.doi.org/10.1093/obo/9780199772810-0268

Lozano, Cristóbal. 2021b. CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. *Second Language Research*, first online view. https://doi.org/10.1177/02676583211050522

Lozano, Cristóbal, Ana Díaz-Negrillo and Marcus Callies. 2020. Designing and compiling a learner corpus of written and spoken narratives: COREFL. In Christiane Bongartz and Jacopo Torregrossa eds. *What's in a Narrative? Variation in Story-telling at the Interface between Language and Literacy*. Bern: Peter Lang, 21–46.

Marques, João. 2013. *Anaphora Resolution in Portuguese: A Hybrid Approach*. Lisbon: Universidade de Lisboa dissertation.

Miltsakaki, Eleni. 2002. Toward an aposynthesis of topic continuity and intrasentential

anaphora. *Computational Linguistics* 28/3: 319–355.
Pardo, Thiago Alexandre Salgueiro, Jorge Baptista, Magali Sanches Duran, Maria das Graças Volpe Nunes, Fernando Antônio Asevedo Nóbrega, Sandra Maria Aluísio, Ariani di Felippo, Eloize Rossi Marques Seno, Raphael Rocha da Silva, Rafael Torres Anchiêta, Henrico Bertini Brum, Márcio de Souza Dias, Rafael de Sousa Oliveira Martins, Erick Galani Maziero, Jackson Wilke da Cruz Souza and Francielle Alves Vargas. 2017. The coreference annotation of the CSTNews corpus. In Raquel Martínez, Julio Gonzalo, Paolo Rosso, Soto Montalvo and Jorge Carrillo-de-Albornoz eds. *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages, vol. 1881*. Murcia: IberEval, 102–112.
Poesio, Massimo and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Daniel Tapias eds. *Proceedings of the Sixth International Conference on Language Resources and Evaluation*. Marrakech: European Language Resources Association (ELRA), 1170–1174.
Pradhan, Sameer, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw and Ralph Weischedel. 2007. Ontonotes: A unified relational semantic representation. In *Proceedings of the International Conference on Semantic Computing*. Irvine, California: IEEE Computer Society, 517–526.
Quesada, Teresa and Cristóbal Lozano. 2020. Which factors determine the choice of referential expressions in L2 English discourse? A multifactorial study from the COREFL corpus. *Studies in Second Language Acquisition* 42/5: 959–986.
Rizzi, Luigi. 1982. *Issues in Italian Syntax*. Dordrecht: Foris.
Siemens, George, Dragan Gaševic and Shane Dawson. 2015. *Preparing for the Digital University: A Review of the History and Current State of Distance, blended, and online learning*. Arlington: Link Research Lab.
Tagnin, Stella. 2006. A multilingual learner corpus in Brazil. In Andrew Wilson, Dawn Archer and Paul Rayson eds. *Corpus Linguistics Around the World*. Brill Rodopi, 195-202.

Corresponding author
Amanda Maraschin Bruscato
Gambelas Campus
8005-139
Faro Portugal
e-mail: a66230@ualg.pt

received: November 2021
accepted: June 2022
published online: July 2022

RiCL Research in Corpus Linguistics

The influence of social prestige on Pino Cacucci's work: A corpus-based study

Virginia Mattioli
Independent researcher / Italy

Abstract – This paper analyses Pino Cacucci's work as a translator and travel writer in order to assess the influence of social prestige on his behaviour when facing otherness. Both translation and travel writing relate different linguistic and cultural contexts to one another. The textual elements representing such linguistic and cultural encounters are foreign words, and their treatment —in terms of maintenance or adaptation— can be used as an indicator of the author's position towards the foreign. From here, the study examines the treatment of foreign words identified in three novels written or translated by Cacucci. Following a corpus-based methodology, the techniques used to transpose foreign words from the source to the target context are determined and related to exoticism (if they maintain the original form) or domestication (if the foreign element is translated or adapted to the target language). Finally, the results are contrasted with the current literary canon. The outcomes reveal a greater acceptance of otherness in the most prestigious novels, in terms of textual practice (translation/travel writing) and linguistic variety (peninsular/Argentinian Spanish), showing the influence of social prestige on the author's behaviour and suggesting some reflections about the relationship between social recognition and acceptance of otherness.

Keywords – Corpus-based translation studies; travel writing; foreign words; translation techniques social prestige

1. INTRODUCTION

This paper assesses the influence of social prestige in Pino Cacucci's work as a translator and travel writer in terms of acceptance of the foreign. Pino Cacucci is an award-winning Italian author, translator and screenwriter specialised in Hispanic language and culture, who travelled extensively across Latin America and lived for long periods in Mexico. Like Cacucci, other authors are involved in travel writing and translation: among many others, Antonio Tabucchi relates the Italian and Portuguese worlds, and Claudio Magris has dedicated himself to German language and culture. Both practices represent contexts in which an encounter with the linguistic and cultural other may occur, and where travellers and translators play the role of mediators, struggling to reconcile the differences arising from contact between the source and the

Research in Corpus Linguistics 11/1: 29–52 (2023). Published online 2022. ISSN 2243–4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.11.01.02

target context. From a linguistic perspective, the elements that better represent such linguistic and cultural encounter are foreign words (Mattioli 2018). As terms adopted from a different language, foreign words highlight the lack of equivalence between the source and the target contexts and become culture-specific representations of the context from which they proceed.

Because of their cultural specificity, the use of foreign words can be considered a sign of the authors' position on otherness. In this sense, a greater use of foreign words shows a greater acceptance of the foreign and the other, tending towards exoticism and foreignisation (Venuti 1995: 20). However, a preference for patrimonial terms by substituting foreign words with elements representing the target cultures suggests a fuller integration of the former; that is, in Venuti's (1995: 18–19) words, "a tendency towards domestication." With this in mind, the present study examines the use of foreign elements in Cacucci's work to observe the degree of acceptance of otherness arising from: 1) the two textual practices in which the author (translation and travel writing), and 2) the linguistic variety (peninsular or Argentinian Spanish). The results will then be contrasted with the Italian literary canon in order to determine its influence on the author's production.

The study is guided by the following research questions:

1. (RQ1) What are the most representative foreign words of the novels under study?
2. (RQ2) What techniques are used to transpose the identified foreign words from the source to the target context in each text?
3. (RQ3) Are the determined techniques related to an exotic or a domestic tendency?
4. (RQ4) What is the author's behaviour in each examined text?
5. (RQ5) Does the author's behaviour change according to the textual practice and the linguistic variety? Are such changes due to the influence of the current literary canon?

The paper is structured as follows. Section 2 deals with the theoretical framework and a review of the literature. Section 3 offers information on the methodology and the corpus of texts analysed, while Section 4 provides a discussion of the results. Finally, Section 5 closes the paper with some final remarks and some proposals for further research.

2. THEORETICAL FRAMEWORK

2.1. Translation and travel

Both translation and travel are intercultural phenomena. On the one hand, translation relates two languages to one another. As language is the primary form of culture (Coseriu 1981: 272–274), translation can be seen as a practice of cultural transfer in which the translator plays the role of cultural mediator (Hatim and Mason 1995: 282). From this perspective, Venuti (1995: 306) defines a translated text as "the site where a different culture emerges, where a reader gets a glimpse of a cultural other." Likewise, to travel allows physical approximation to different cultures and languages, a characteristic that is reflected in travel literature: the literary product of a travel experience. According to Díaz Larios (2007: 127), the main objective of travel literature is the cultural translation from the source to the target context by means of expressions and terms proper to the language of the visited country, accompanied by translations and explications.

Following the above-mentioned function of cultural mediation, some authors relate translation and travel to one another, identifying and describing their similarities. A frequent comparison is made between translation and migration. Carbonell i Cortés (2003: 145) proposes such a comparison, referring to the transference of the text from the source to the target language and culture as a proper migration. A similar position is held by Trivedi (2005), who defines translation as "a process and condition of human migrancy." Other authors (cf. Baynat Monreal 2007) relate practices of translation and travel to a sort of exploration, from a linguistic or geographic perspective respectively. Others compare the textual products originated by translation and travel in terms of common aims. Both practices present a global reach translation by promoting the global circulation of knowledge, and travel literature as a real global literary type that describes the main social, cultural and economic phenomena of the modern era (Pickford and Martin 2013: 2). Similarly, Ortega Román (2006: 221–223) underpins the linguistic and informative objective of travel writing, which, as translation, often includes terms and expressions of the visited country, offering readers lexical, semantic and etymological information about the language spoken there.

In addition to having a similar nature and aim, translation and travel writing also share some specific characteristics. Firstly, both practices lead to the discovery of the otherness (Smecca 2003; Carbonell i Cortés 2014) by relating different contexts and

cultures to one another. Secondly, both translators and travel authors occupy a marginal role within society, in contrast to the dominant thought. From a socio-political perspective, a translator plays the role of a social agent who can resist economic and political power (Osinaga 1999: 378–379). Thus, translators can be used by institutions to impose certain cultures or values or, conversely, subvert the dominant ideology (Xianbin 2007). Also, from a linguistic perspective, both practices are characterised by the interference of linguistic and cultural codes (Polezzi 2012) from which an inherent plurilingualism originates, often considered a form of deviance with respect to the less common but normative monolingualism (Tymoczko 2006: 16).

The third feature shared by translation and travel writing is the 'in-between' state that characterises both practices from several perspectives: 1) physically, considering travel authors' and translators' liminal position between the source and target language, and culture (Mattioli 2018: 90); 2) linguistically, due to the inherent interference between different codes (Díaz Larios 2007: 127–128; Bennet 2012: 8; see Section 2.2); and 3) culturally, for the presence of foreign words that reveal the distance between cultures and contribute to bring them closer (Bhabha 1994: 325–326). Such a liminal state and knowledge of both linguistic and cultural codes can lead translators and travellers to act as cultural mediators, trying to reconcile and bring closer the source and target context using similar techniques. One of the most powerful techniques used in both practices to represent the encounter with the other is the use of foreign words.

2.2. Foreign words

As hybrid formats, both translation and travel writing tend to use foreign words and expressions. Since the 1970s, scholars in Translation Studies who approach the discipline from feminist and postcolonial perspectives consider the interference of foreign languages an inherent characteristic of translation, whose production assumes the denomination of "in-between language" (Mattioli 2018: 90). In the present century, Bennet (2012: 8) adopts a different angle on the same concept and describes translation linguistic hybridity with the expression "writing-as-translation." Foreign words are essential elements also in travel writing, allowing the author to transpose the culture of the visited country (Díaz Larios 2007: 127–128) and describe novel realities by adding "local colour" and exoticism to the text (Curell *et al.* 2010: 49).

There is still no clear consensus as regards the definition of foreign words which are commonly explained by resorting to various terms (loans, borrowings, foreign words, transpositions, etc.), depending on their origin and/or degree of acceptance in the target language. In the present study, the term 'foreign word' is used in its most general sense and relates to the definition of the term 'extranjerismo' in the *Diccionario de la Real Academia Española* (DRAE; Real Academia de la Lengua 2001): "Voz, frase o giro que un idioma toma de otro extranjero" ('Any voice, phrase or expression that a language adopts from a foreign one'). Such a definition includes all the terms used in a (target) language that proceed from a different (source) language, regardless of their degree of adaptation. Consequently, all terms with any foreign linguistic feature will be considered foreign words, whereas the labels 'loan', 'borrowing', 'transposition', etc. will be used to define specific techniques chosen by translators to transpose foreign words from a source to a target language (see Section 3.2). The definition in the DRAE, on the one hand, allows for a systematic recognition of foreign words in a text, by considering all those terms in which linguistic features do not respect the word formation rule of the target language and, on the other, it emphasises the cultural character of foreign words, supporting the possibility of using them as indicators of the acceptance of otherness. By focusing on their origin in a different source language, the adopted definition underlines the word's foreign origin and highlights the process of cultural transference needed to transpose it to a different target context. As shown by Mattioli (2018: 56), this foreign origin is precisely what makes foreign words representations of otherness.

However, maintenance of foreign words is only one of the techniques employed by authors and translators to reconcile the cultural differences between the source and target context. Often, to overcome the semantic opacity of foreign elements, other techniques are used, such as combining them with explanations or replacing them with other linguistic elements (Curell *et al.* 2010: 49). The determination and classification of the techniques used to transpose cultural elements has attracted the attention of many authors, particularly in Translation Studies. Among the main proposals, some scholars (Nida 1964; Newmark 1988; Molina Martínez 2006) classify such techniques into discrete categories. The resulting taxonomies can include fewer, more general classes of translation techniques or many, more specific classes. Among the former, the taxonomy proposed by Nida (1964: 226–239) organises "techniques of adjustment" (Nida 1964:

226) into five classes: additions, subtractions, alterations, use of footnotes and adjustments of language to experience. Among the latter, Molina Martínez (2006: 101–104) distinguishes 18 different classes: adaptation, linguistic ampliation, amplification, calque, compensation, linguistic compression, discursive creation, description, recognised translation, generalisation, modulation, particularisation, loan, reduction, substitution, literal translation, transposition and variation.

Other authors (Hervey and Higgins 1992; Mangiron 2006; Carbonell i Cortés 2014) consider translation techniques on a continuum, from those that maintain the original culture (that is, exoticism) as much as possible to those that adapt to the target culture (that is, domestication or adaptation) as much as possible. In the present study, a specific classification has been adopted, combining both types of taxonomies (see Section 3.2).

2.3. Polysystem studies and social prestige

The last topic to review for the purpose of analysis is the literary system and the relationships between different genres and languages, in order to observe Pino Cacucci's behaviour regarding the acceptance of otherness in relation to the two variables considered in this study: the prestige of the textual practice (translation/travel writing) and linguistic variety (peninsular/Argentinian Spanish). According to Even-Zohar (1990), any culture is represented by a polysystem that includes as many systems as cultural fields (literature, anthropology, sociology, translation, etc.). The polysystems representing each culture (Spanish, English, Italian, etc.) are then included within a global polysystem. Within the global polysystem, each cultural polysystem dynamically correlates with others and occupies a central or peripheral position according to its degree of canonisation and standardisation. The canon is decided by the dominant class. The same happens with systems within each cultural polysystem. As a result, the central position of each (poly)system is occupied by the most accepted, canonical and prestigious cultures —hence languages— or cultural fields, while the periphery holds the less prestigious.

Within the current literary system, the canon is represented by the monolingual literature written in the most prestigious language. Consequently, both translation and travel writing, as practices that combine different codes in the same texts (Polezzi

2012), occupy a peripheral position. However, not all translated works occupy the same position since, according to the norms of the polysystem, within the translation system, translated texts hold different degrees of prestige. The recognition of translated texts depends on the centrality of the original work, its original and target languages, and its literary genre. The novels analysed in this study are Spanish novels translated into Italian. Both languages are peripheral from a geo-political perspective (Mattioli 2018: 66); however, peninsular Spanish is the variety with the greatest prestige (Lope Blanch 1972; Slebus 2012: 29), holding a more central position than Argentinian Spanish. As for the genre, the original versions of the novels that have been selected for analysis have received significant international literary prizes, occupying a very central position in the literary polysystem. Consequently, the corresponding translations also boast a central position, even if not as central as the original texts. On the contrary, travel novels, as belonging to the genre of travel writing, occupy a peripheral position determined by their authors' status and their hybrid style. It should be borne in mind that, on the one hand, travel writing is often considered a "minority discourse" (Bhabha 1994: 325) for the migrant condition of travel authors, who are at a distance from the dominant (hence canonical) thought and assume a liminal position that tends to overcome the limits and the impositions of the proper cultural context to include different cultures encountered during the travel experience (Mattioli 2018: 88). On the other hand, migrants' textual production is described by many scholars as a "polylingual writing" (Polezzi 2012: 351) for its inherent hybridity, expressed in the interference of different codes. Such hybridity contrasts with the prestige of monolingual discourse and, according to Bakhtin (1981 [as cited in Polezzi 2012: 351]), as a kind of polyglossia (that is, plurilingualism within the same linguistic community), it resists the dominant tendency to centralisation and control by disintegrating the unity, as opposed to monoglossia (that is, monolingualism), which fosters centralisation, instead.

Each pair of novels representing the two variables analysed in this research (textual practice and linguistic varieties) presents different degrees of social prestige: the translated novel selected for analysis occupies a more central position than the travel novel, whereas the novel translated from peninsular Spanish holds a more central position than that translated from Argentinian Spanish.

3. SOURCES AND METHODOLOGY

3.1. Sources

The analysed material consists of three novels, written in Italian and published between 2000 and 2015, which represent Cacucci's travel writing and translation activity: 1) *Le Balene lo Sanno* (Cacucci 2009), an Italian travel novel about one of the author's journeys to Mexico; 2) *Soldati di Salamina* (Cercas 2002), translated into Italian by Cacucci from the original Spanish novel *Soldados de Salamina* (Cercas 2001); and 3) *Bersaglio Notturno* (Piglia 2011), translated into Italian by Cacucci from the original Argentinian novel *Blanco Nocturno* (Piglia 2010). Table 1 shows the number of types and tokens of each novel which make up of the corpus examined in the study:

Novel	Types	Tokens
Le Balene lo Sanno	9, 506	159,466
Soldati di Salamina	9,563	242,924
Bersaglio Notturno	11,640	305,348
Total	-	707,738

Table 1: Length of the novels under examination

These novels were chosen because they have been awarded with a prestigious, international, literary prize conferred in Italy or in Spain, hence being recognised as representative of their genre according to the literary canon.[1] The comparison of the results across the novels allows to observe the influence of social prestige on the author's approach to otherness, according to the textual practice (translation or travel writing) and the linguistic variety (peninsular or Argentinian Spanish).

3.2. Methodology

The corpus-based methodology adopted for the present research consists of five steps, which corresponds to the five research questions that guide the study:

1. Firstly, foreign words identified in Mattioli (2018) were selected and adopted as a point of departure for the present research.
2. Secondly, the techniques used for the transposition of these foreign words from the source to the target context were determined.

[1] *Le Balene lo Sanno* won the *Premio Salgari* in 2010, a prestigious Italian literary award dedicated to travel literature. *Soldati di Salamina* won the Italian international literary prize *Premio Grinzane Cavour* for the section of 'Foreign fiction' in 2003. *Blanco Nocturno* won the 2011 Spanish *Premio de Novelas Rómulo Gallego.*

3. Then, each determined technique was related to an exotic or domestic tendency.
4. Next, the obtained results were compared across the texts analysed.
5. Finally, the results obtained were contrasted with the current literary canon.

In what follows, each methodological step is described in detail. As regards the first step, the study is based on Mattioli (2018) in that it takes as point of departure its most representative foreign words that were retrieved from a set of 47 novels with the use of electronic corpus-based methods. As in Mattioli (2018), the present study considers foreign words to include all those terms presenting linguistic features different from those allowed by the Italian word formation rule and assesses their representativeness according to qualitative and quantitative criteria. As far as qualitative representativeness is concerned, only the foreign words of three semantic fields are examined: 1) food and drink, 2) clothing and bodily care and 3) communication and transportation. Such a choice allows, on the one hand, to include a minor number of elements and therefore be able to analyse them more accurately and, on the other, to consider only those foreign words strictly related to cultural contact. For instance, food and clothing are not only extremely culture-specific but represent the primary necessities of any population and are attested in any culture. Similarly, transportation and communication make cultural contact possible by allowing physical displacement and the dissemination of ideas and concepts.

From a quantitative perspective, Mattioli (2018) selects only those foreign words, representing one of the three chosen semantic fields, which present a total frequency equal to or higher than ten occurrences in the entire corpus, and which are present in at least three of the 47 novels analysed. This allows Mattioli to examine only those foreign words which actually represent the textual practice under analysis (that is, translated or travel novels) and, at the same time, discard those items which are used only in a specific novel or are related to a specific geographical area. For the present study, only the items identified by Mattioli (2018) in the three texts under study are considered. As a result, not all the foreign words in the examined novels are analysed, but only those meeting the representativeness criteria considered in Mattioli (2018).

Once the foreign words are retrieved, each of them is examined separately to identify the technique used to transpose them from the source to the target context. A translation technique taxonomy is established by combining the two most common

types of classification proposed by previous scholars in the field of Translation Studies: 1) the classification of translations techniques into discrete categories, and 2) the classification which orders the translation techniques in a continuum from the most exotic to the most domestic (see Section 2.2). This moves us to step three in which a selected set of discrete categories is ordered in a continuum, from the most exotic to the most domestic.

Tendency	Technique	Description	Example
Exoticism	**Addition**	Addition of an explanatory expression (parenthesis, paraphrase, etc.)	*They expected him to dismount* > *aspettando che scendesse dalla Jeep* ('waiting for him to dismount from the Jeep')
	Transposition	Foreign item not included in the TL dictionary	*Parkway* > *Parkway*
	Loan/Borrowing	Foreign item included in the TL dictionary as a loan	*Jeans* > *Jeans* *Jeans* ‹ǧìinʃ› o ‹ǧèinʃ› s. ingl. [pl. di jean, 'tipo di tela che prob. trae il nome dalla città di Genova (v. blue jeans)], usato in ital. al masch.' (sing. o pl.). (*jeans* ‹ǧìinʃ› o ‹ǧèinʃ› English noun [pl. of jean, a kind of weave that probably takes its noun from the city of Genova (see blue jeans)], used in Italian in masculine (sing. or pl.). (*Treccani Online Dictionary of the Italian language*, my translation)[2]
	Naturalisation	Item adapted to the target language phonetics and/or phonology	*Pullman* > *Pulman*
	Modulation	Use of a more general/more specific term	*The lorry* > *il mezzo* (the mean)
	Literal translation	Translation provided by the bilingual dictionary	*Sandwich* > *Panino* (sandwich)
	Omission	Elimination of the foreign word	*Watching a western film* > *guardare un Ø western* ('watching a western Ø')
	Lack of equivalence	Use of a translation not equivalent to the original text	*I had our computer tested* > *ho fatto controllare il sistema di sicurezza* ('I had our security system tested')
Domestication/ Adaptation			

Table 2: Translation technique taxonomy adopted in the analysis

This is illustrated in Table 2 above, where each technique is presented, together with its description and an example of use (for each Italian example, an English translation is

[2] https://www.treccani.it/vocabolario/jeans/

provided for a better understanding). Next, the translation techniques taxonomy is modified for application to the travel novel which, as an original text, cannot be analysed by means of a parallel corpus. Therefore, the techniques 'omission' and 'lack of equivalence' are considered as non-applicable.

To determine which technique is used to transpose each foreign word from source to the target context, each occurrence is observed and related to the corresponding technique. This task differs for the translated and the travel novel. For translated novels, a parallel corpus including the original and the translated version of each novel is analysed by using the parallel corpus concordance *AntPConc* (Anthony 2017). By searching for each foreign word in the translated corpus it is possible to view it in context, both in the translated and the corresponding original text, allowing for comparison. The search is then repeated, starting from the original corpus, to detect cases of translation, modulation, adaptation, omission and lack of equivalence in which the translator has omitted the foreign word or has replaced it with a patrimonial term, a paraphrasis or any other lexical item. Observing and comparing each pair of the occurrences retrieved (original and translated), the translation technique used in each case is determined.

A different method is, however, used to examine the travel novel, as the text proceeds from an experience rather than from another original text. With the use of *AntConc* (Anthony 2020), each foreign word is searched for in the concordance list to observe it in context. Then, by comparing the use of each term with the adopted taxonomy and with the help of Italian monolingual and English-Italian bilingual dictionaries, the technique used in each case is determined. However, it must be borne in mind that this method allows only the determination of techniques tending towards exoticism, in which foreign words are maintained (transpositions, loans and naturalisations). Consequently, in order to detect those cases in which further techniques are used to transpose the term from the source to the target context (translation, modulation and adaptation), the synonyms of each foreign word (both patrimonial and foreign words accepted in Italian as loans) are retrieved and looked up in the concordance list. In case of a foreign word already analysed in the corpus of translated novels, the resulting terms are considered synonyms; in case of foreign words identified exclusively in the travel novel, its synonyms are sought in the Italian thesaurus. An

example of this process, which is focused on the foreign word *calle* (Spanish 'street'), is provided in Table 3.

Task	**Results**
Search for *calle* in the concordance list	*casa in calle Topete* ('house in calle Topete')
Search for *calle* in the Italian dictionary	No result
Determination of the technique used among those included in the adopted taxonomy (according to the description provided in the Table 2)	Transposition (foreign item not included in the target language dictionary)
Search for the word *calle* among the foreign words identified in the translated novels	Previously identified
Use of the results obtained from the analysis of the translated novels to retrieve synonyms (loans and patrimonial words) of the word *calle*	*via* ('street') *strada* ('road') *stradina* ('small road')
Search for each retrieved synonym in the concordance list: *via*, *strada*, *stradina*	*Una via del centro* ('a street in a downtown area') *Ai lati della strada* ('on the borders of the road') *Stradina*: no results
Determination of the technique used among those included in the adopted taxonomy (according to the description provided in the Table 2)	*Una via del centro* ('a street in a downtown area') > literal translation (translation provided by the bilingual dictionary) *Ai lati della strada* ('on the borders of the road') > literal translation (translation provided by the bilingual dictionary)

Table 3: Example of the process to determine the transposition techniques of the foreign words in the travel novel

Once the techniques used in each case to transpose the foreign words under examination from the source to the target language have been determined, they are related to an exotic or domestic tendency by associating each technique with a specific type of behaviour: addition, maintenance, adaptation, translation or omission. Table 4 shows the distribution of the techniques among the five types of behaviours.

Technique	**Behaviour**	**Tendency**
Addition	Addition	Exoticism
Transposition	Maintenance	
Loan/Borrowing		
Neutralisation	Adaptation	↕
Modulation		
Literal translation	Translation	
Omission	Omission	
Lack of equivalence		Domestication/Adaptation

Table 4: Relation between techniques, behaviours and tendency towards exoticism or domestication

Then, the tendency towards exoticism or domestication of each behaviour is determined according to their position on the continuum. Finally, for each novel, the total number of occurrences of each technique and of each behaviour are calculated.

In the fourth step, the results yielded by the analysis are compared across the three examined novels in quantitative terms, namely, comparing the number of foreign words identified in each novel, the number of occurrences of each technique and of each behaviour. To do so, the three novels are contrasted as follows: firstly, translated novels are contrasted with travel novel and, secondly, the novel translated from peninsular Spanish is contrasted with the novel translated from Argentinian Spanish. The novels are of different weight in terms of tokens. Consequently, statistical difference in the tokens is considered when comparing the results across the different texts. To calculate this, a Log-likelihood (LL) test is used considering as significant only results equal to or higher than the threshold 6.63, that is, accepting a *p* value lower than 0.01.

Finally, the outcomes of the comparisons in step four are contrasted with the position occupied by the novels within the literary polysystem (See Section 2.3). This allows to verify whether Cacucci's treatment of foreign words can be related to the influence of the current literary canon and the corresponding social prestige. Hence, the results of the comparisons between the two textual practices (translation or travel writing) and between the two linguistic varieties (peninsular and Argentinian Spanish), as far as the treatment of the foreign words is concerned, are contrasted with the position occupied by each novel within the literary polysystem, in order to detect any correlation.

4. RESULTS AND DISCUSSION

The results yielded by the corpus-based analysis seem to demonstrate that the literary canon and consequent social prestige influence the author's behaviour when encountering foreign words and, thus, his degree of acceptance of otherness and of cultural differences. Specifically, the results show that the more canonical the genre or the linguistic variety, the greater the acceptance of otherness as arising from the use of foreign words. Table 5 shows the most representative foreign words identified and their number of occurrences in each of the three texts analysed in Mattioli (2018), which is used as a departure point in this case study. It is worth mentioning, however, that some

of the foreign words analysed do not occur in the novels of the present study in their original foreign form. Such words are identified in other novels included in the larger corpus examined in Mattioli (2018) and, according to the representativeness criteria adopted in her study, have been examined as particularly representative. According to Mattioli (2018), such items do appear in the novels under study in adapted or translated forms. Consequently, considering their representativeness, such foreign words have been included in the present analysis, as they allow for exploring the author's behaviours beyond transpositions or loans.

Le Balene lo Sanno		***Soldati di Salamina***		***Bersaglio Notturno***	
Foreign word	**Frequency**	**Foreign word**	**Frequency**	**Foreign word**	**Frequency**
Autobus	0	*Autobus*	1	*Autobus*	3
Bike	0	*Avenida*	0	*Calle*	4
Calle	1	*Bistrot*	8	*Camion*	9
Carretera	8	*Calle*	6	*Gilet*	1
Camion	3	*Cognac*	4	*Gin*	5
Canoa	2	*Computer*	5	*Jeans*	3
Computer	2	*Gin*	0	*Reportage*	1
Email/e-mail/mail	2	*Jeans*	1		
Film	12	*Whisky*	4		
Jeep	0				
Pick-up	3				
Poncho	0				
Sombrero	1				
Tunnel	2				
Yucca	3				

Table 5: The most representative foreign words analysed by Mattioli (2018) in the three semantic fields in the three novels

All the occurrences of each foreign word have been examined and contrasted with the corresponding source or patrimonial terms. As a result, the total number of items under examination has increased, because of the addition of adaptations and patrimonial terms to the foreign words. Adaptations and patrimonial words have been added to the amount of foreign or patrimonial words depending on their respect for the Italian word formation rule (see Section 3.2). The total number of items analysed is shown in Table 6, contrasting the number of foreign and patrimonial elements for each novel in terms of number of occurrences and percentage.

Novels	**Foreign words**		**Patrimonial terms**		**Total**
	Tokens	**Percentage**	**Tokens**	**Percentage**	
Le Balene lo Sanno	39	41%	55	59%	**94**
Soldati di Salamina	39	71%	16	29%	**55**
Bersaglio Notturno	26	38%	42	62%	**68**

Table 6: Total items (foreign and patrimonial terms) analysed in each novel

The data show that, in the travel novel, the total number of items that are analysed is greater than in the translated texts. Both numbers are statistically significant with a LL of 33.3 contrasted with the Spanish novel, and of 39.5 against the Argentinian novel. However, the amount of foreign items with respect to patrimonial items is greater in the novel translated from the original Spanish text (71%) than in the travel novel (41%), and a very small difference arises from the comparison between the travel novel and the novel translated from the Argentinian original (38%). Such primary results suggest a more frequent use of foreign words in the translated novels than in the travel text.

The preference for foreign or patrimonial terms also varies according to the source language variety. Although in both translated novels a similar amount of items is analysed (55 items in the Spanish novel and 68 in the Argentinian, with no statistical difference: LL: 0.04), the percentage of items maintained in their original form is higher in the novel translated from peninsular Spanish than in that translated from Argentinian Spanish. These primary outcomes are further underpinned by the results obtained from the determination of the techniques used for the transposition of identified foreign words.

As is the case with the proportion of foreign words, the techniques used to transpose them from the source to the target context also change depending on the textual practice and the linguistic variety. The most frequent choice in the travel novel examined is a translation (55 cases; corresponding to 59% of the total instances examined), followed by maintenance (36 cases; 38%) and adaptation (3 cases; 3%). According to the positions occupied by such behaviours within the continuum extending from exoticism to domestication, the author tends predominantly towards domestication, and prefers to translate the foreign elements, hence adapting them to the target culture. Similar outcomes arise from the analysis of the novel translated from Argentinian Spanish in which, on 38 out of 68 times (56% of the total occurrences in the novel), the author translates the foreign elements, in four cases (6%) he omits them and in 26 (38%) he maintains them in their original form. Again, these data show a tendency towards domestication in translating from Argentinian Spanish.

The outcomes change consistently for the novel translated from peninsular Spanish. Here, foreign elements are maintained 71 per cent of the time, corresponding to 39 occurrences out of 55 analysed, and translated just in 16 cases (29% of the total number of instances examined). In this case, the translator’s behaviour presents a

predominant tendency towards exoticism, maintaining the original, foreign forms, and remaining faithful to the source culture. The data retrieved from the examination of the techniques used to transpose foreign words from the source to the target context in each novel, their corresponding behaviours and their tendency within the continuum are shown in Table 7.

Le Balene lo Sanno						
	Techniques			**Behaviours**		**Tendency**
Technique	**Tokens**	**Percentage**	**Behaviour**	**Tokens**	**Percentage**	
Addition	0	0%	Addition	0	0%	Exoticism
Transposition	9	10%	Maintenance	36	38%	
Loan/Borrowing	27	29%				
Neutralisation	3	3%	Adaptation	3	3%	
Modulation	0	0%				
Literal translation	55	59%	Translation	55	59%	Domestication
Total	94	100%	**Total**	94	100%	
Soldati di Salamina						
	Techniques			**Behaviours**		**Tendency**
Technique	**Tokens**	**Percentage**	**Behaviour**	**Tokens**	**Percentage**	
Addition	0	0%	Addition	0	0%	Exoticism
Transposition	15	27%	Maintenance	39	71%	
Loan/Borrowing	24	43%				
Neutralisation	0	0%	Adaptation	0	0%	
Modulation	0	0%				
Literal translation	16	29%	Translation	16	29%	
Omission	0	0%	Omission	0	0%	Domestication
Lack of equivalence	0	0%				
Total	55	100%	**Total**	55	100%	
Bersaglio Notturno						
	Techniques			**Behaviours**		**Tendency**
Technique	**Tokens**	**Percentage**	**Behaviour**	**Tokens**	**Percentage**	
Addition	0	0%	Addition	0	0%	Exoticism
Transposition	4	6%	Maintenance	26	38%	
Loan/Borrowing	22	32%				
Neutralisation	0	0%	Adaptation	0	0%	
Modulation	0	0%				
Literal translation	38	56%	Translation	38	56%	
Omission	3	4%	Omission	4	6%	Domestication
Lack of equivalence	1	1%				
Total	68	100%	**Total**	68	100%	

Table 7: Techniques used in each novel to transpose the foreign words from the source to the target context

The data further show that, in *Bersaglio Notturno*, maintenance is found in 22 cases (32% of the total occurrences) by using a loan and in four instances (6%) by means of transposition. By contrast, *Soldati di Salamina*, exhibits a more frequent use of transpositions (27%, representing 15 cases out of 55) and *vis-à-vis* loans (43%; 24 occurrences). As pointed out in Section 3.2 (cf. Table 2), transposition includes foreign

terms not included in the target language dictionary —in this case, Italian— whereas loans come from a foreign language but have been already accepted in the target language, and thus included in the dictionary. Consequently, transpositions represent a more exotic and distant otherness than loans and their more frequent use is a further sign of the acceptance of otherness in the novel translated from peninsular Spanish as opposed to that translated from Argentinian Spanish.

Finally, in each novel the proceeding language of the foreign words being examined has been considered. Previous literature distinguishes between the original language of the foreign words and that from which they were adopted, even if they previously originated in a different language (Degerstedt 2013). For example, the word *yuca* ('yucca') identified in the travel novel is from Maya; however, it was introduced into Italian from Spanish. In this case study, the language of introduction that has been considered is the most relevant one for the study, that is, the language related to the cultural contact. Table 8 shows the proceeding languages of the foreign words identified in each text.

Novel	Foreign word	Source	Frequency foreign form	Frequency translated form
Le Balene lo Sanno	*Calle*	Spanish	1	22
	Carretera	Spanish	8	1
	Canoa	Spanish	2	0
	Poncho	Spanish	0	1
	Sombrero	Spanish	1	0
	Yucca	Spanish	3	0
	Total Spanish tokens		**15**	**24**
	Bike	English	0	1
	Computer	English	2	0
	Email/e-mail/mail	English	2	1
	Film	English	12	0
	Jeep	English	0	5
	Pick-up	English	3	0
	Total English tokens		**19**	**7**
	Autobus	French	0	1
	Camion	French	3	0
	Tunnel	French	2	23
	Total French tokens		**5**	**24**
	Total tokens		**39**	**55**

Table 8: Source languages of the foreign words identified in the corpus

Novel	Foreign word	Source	Frequency foreign form	Frequency translated form
Soldati di Salamina	*Computer*	English	5	0
	Gin	English	9	0
	Jeans	English	1	0
	Whisky	English	4	0
	Total English tokens		**19**	**0**
	Autobus	French	1	4
	Bistrot	French	8	0
	Cognac	French	4	0
	Total French tokens		**13**	**4**
	Avenida	Spanish	0	1
	Calle	Spanish	7	11
	Total Spanish tokens		**7**	**12**
	Total tokens		**39**	**16**
Bersaglio Notturno	*Autobus*	French	3	2
	Camion	French	9	0
	Gilet	French	1	0
	Reportage	French	1	0
	Total French tokens		**14**	**2**
	Gin	English	5	0
	Jeans	English	3	0
	Total English tokens		**8**	**0**
	Calle	Spanish	4	40
	Total Spanish tokens		**4**	**40**
	Total tokens		**26**	**42**

Table 8: (continuation)

When comparing the travel novel to the translations, the data show that only the former presents a balanced number of foreign words from English and from Spanish, the language of the visited country, with six types representing each proceeding language. On the contrary, in both translated novels, the foreign words from Spanish are very restricted in terms of type, even if they come from the source language of the original text: two types in *Soldati di Salamina* and just one in *Bersaglio Notturno*.

As regards the relationship between the proceeding language of the foreign item and its maintenance or transposition into the target language, in the travel novel 15 out of the 39 occurrences of Spanish words (38% of times) are maintained in their original form, being the second most frequently maintained only surpassed by English (with 19 out of 26 occurrences, 73% of times). By contrast, in both translated novels, Spanish words are usually translated. Here, the cases of maintenance are seven out of 19, corresponding to 39 per cent in the novel translated from peninsular Spanish, and four out of 44, representing 9 per cent in the novel translated from Argentinian Spanish.

As for the comparison between the two Spanish varieties, in *Soldati di Salamina* foreign words mostly come from English (four out of nine types), whereas in *Bersaglio*

Notturno they come mainly from French (four out of seven types). Regarding the relationship between the author's behaviour and the proceeding language of the foreign words, when translating from peninsular Spanish, Cacucci tends to maintain English words in their original form in all cases, French words in 13 cases out of 17 (76% of times) and Spanish terms in six out of 19 (39% of times). The author's adoption of different behaviours depending on the proceeding language of the foreign words is even more noticeable in the novel translated from Argentinian Spanish. Here, Cacucci maintains English words in their original form in all cases and French words in 14 out of 16 cases (87% of times), whereas he translates Spanish terms, which are maintained only in 9 per cent of cases (four occurrences out of 44).

In the last stage of the analysis (see Section 3.2), the results obtained from the comparison across the novels examined here are contrasted with the literary canon. According to the current canon, monolingualism is more commonly accepted than multilingualism and, in the case of the awarded-winning novels that have been examined, the textual practice of translation is more prestigious than travel writing (see Section 2.3). Contrasting this with the results obtained from the comparison of the travel novel with the novel translated from peninsular Spanish allows to assess the influence of social prestige on the author's behaviour according to textual practice (translation vs. travel writing). The data show that when translating a prestigious novel from the most accepted variety of Spanish, Cacucci uses more exotic techniques by including a greater number of terms in their original form. Such a behaviour, on the one hand, reveals a greater acceptance of otherness and the foreign and, on the other, represents a deviance from the canon as the preference for foreign elements vis-à-vis patrimonial ones gives rise to a hybrid code and style that tends towards multilingualism. Meanwhile, in travel writing —more peripheral in the literary system, hence less socially recognised— the author follows the canon by substituting foreign words with patrimonial terms by means of techniques that tend towards domestication and adaptation of the foreign culture to the target culture.

The same differences arise from the comparison between the novels translated from the two Spanish varieties. Here, Cacucci draws away from the canon by introducing foreign elements and by showing a greater acceptance of otherness in the translation from peninsular Spanish, the more prestigious linguistic variety. On the contrary, when translating from Argentinian Spanish, the author demonstrates a greater

acceptance of the literary canon and integrates otherness in the target culture by means of translations and adaptations of the foreign elements, in order to avoid hybridity and multilingualism. An exception to this, however, can be attested when considering the proceeding language of the examined foreign words. In the three novels, the terms that are analysed proceed from the same three languages: English, Spanish and French. For its capability to integrate foreign elements (Díaz Prieto 1998: 167), for the preference for its use as global language of communication (Gheno 2019: 462) and for the social prestige of Anglophone countries and cultures (Grochowska 2010: 48), English occupies a very prominent position within the linguistic and literary system. French, in turn, was the language of culture and prestige until the end of the previous century, and hence occupies a less central position in the linguistic and literary system. Spanish is the most peripheral language of the three in terms of acceptance and social consideration since the eighteenth century, when, after the golden period of the Spanish language, its prestige decreased to the detriment of French, the globally recognised language of culture for the next century (Porras Castro 1999: 612).

In both translated novels, the author seems to follow the literary canon by opting for words proceeding from the more central languages, namely, English and French. By contrast, in the travel text, preference is given to foreign words proceeding from Spanish, the most peripheral of the three foreign languages but the most representative of the source culture. Cacucci's choices seem to highlight the role of cultural representation of foreign words in the travel novel; that is, even if, according to the literary canon, foreign elements are usually translated or adapted to the target language, when maintained, they are used to represent the source culture.

5. Final remarks

This paper has assessed the influence of social prestige on Pino Cacucci's choices when facing otherness as a translator and a travel writer based on his treatment of foreign words in two textual practices. Translation and travel writing are comparable in cultural terms, representing contexts in which an encounter with the other takes place. From a textual perspective, such an encounter is represented by foreign words, which are evidence of cultural and linguistic contact and, as a textual representation of the foreign, can be used as an indicator of the author's degree of acceptance or rejection of otherness: the greater their use, the greater the acceptance of the foreign.

The corpus of texts has allowed for the examination of the authors' behaviour facing foreign words —thus, otherness— considering two textual practices (translation and travel writing) and two Spanish varieties (peninsular and Argentinian) with different degrees of social prestige.

The results of the five-step analysis described in Section 3.2 suggest that the author's respect for canonical literary norms varies according to the textual practice and the source language variety, showing the influence of social prestige on his decisions when translating. Considering the use of foreign words and the tendency to adopt a more exotic and hybrid style which tends towards multilingualism, Pino Cacucci follows the canon in writing travel texts and in translating from the more peripheral Argentinian Spanish but draws away from accepted literary norms when translating famous and accepted novels from the most prestigious Spanish variety. Such differences prompt two reflections. On the one hand, the tendency towards exoticism in the more recognised practice and in translating from the more normative linguistic variety could be considered a conscious or unconscious attempt to break away from the norms of the literary canon. On the other, the tendency towards domestication and the preference for patrimonial terms in the less prestigious practice and in translating from a more marginal linguistic variety may suggest a search for greater acceptance and centrality within the literary system.

Further, the results prompt reflections on the relation between social recognition and otherness. This study has considered the use of foreign words as evidence for greater acceptance of otherness and a tendency towards domestication as a textual representation of the greater integration of the otherness within the target culture. Combining this assumption with canonical literary norms related to the use of foreign elements studied for the present research (see Section 2.3) may presume the existence of an interrelation between the degree of acceptance of others and social recognition. If a preference for foreign words is related to a greater acceptance of otherness, and their introduction in the literary texts conveys less social recognition within the literary system, then the acceptance of otherness also conveys less social recognition. Similarly, as the preference for adapting foreign elements and patrimonial terms is related to a greater integration of otherness in the target culture and enjoys greater social prestige, then the tendency to integrate otherness into one's own culture also enjoys greater social prestige. From this perspective, Pino Cacucci's effort to break the current canonical

norms by introducing foreign words —hence creating a hybrid, multilinguistic style— in the most accepted and central novel analysed (*Soldati Di Salamina*) can be seen as an attempt to subvert the established social norms that determine the canon, in order to reconcile social recognition with a greater acceptance of otherness.

The present study can be used as a departure point for several further research from both methodological and conceptual perspectives. The methodology that has been adopted here could be replicated or/and improved to analyse other types of foreign words. It would be worth, for example, investigating different semantic classes or focusing only on terms proceeding from a certain original language. Furthermore, the methodology could be fruitfully used to examine the transposition of other features from the source to the target context in any type of corpus, either parallel or comparable, which would allow to compare translation with other types of textual practices. From a conceptual prism, the results obtained in this investigation can be broadened in many senses, for example, by analysing further features related to social prestige and the literary canon or by exploring further policies that can be adopted facing the current canonical norms. This, undoubtedly, represents an avenue for further research.

REFERENCES

Anthony, Lawrence. 2017. *AntPConc* (version 1.2.1). Tokyo, Japan: Waseda University. https://laurenceanthony.net/software

Anthony, Lawrence. 2020. *AntConc* (version 3.5.9). Tokyo, Japan: Waseda University. https://laurenceanthony.net/software

Bakhtin, Mikhail Mikhailovich. 1981. *The Dialogic Imagination: Four Essays*. Austin: University of Texas Press.

Baynat Monreal, María Elena. 2007. El traductor de relato de viajes: De París a Cádiz de Dumas. In Francisco Lafarga, Pedro S. Méndez Robles and Alfonso Saura Sánchez eds., 63–73.

Bennet, Karen. 2012. At the selvedges of discourse: Negotiating the "In-Between" in Translation Studies. *Word and Text: A Journal of Literary Studies and Linguistics* II/2: 43–61.

Bhabha, Homi K. 1994. *The Location of Culture*. London: Routledge.

Cacucci, Pino. 2009. *Le Balene lo Sanno*. Milano: Feltrinelli.

Carbonell i Cortés, Ovidi. 2003. Semiotic alteration in translation: Othering, stereotyping and hybridization in contemporary translations from Arabic into Spanish and Catalan. *Linguistica Antverpiensia* 2: 145–159.

Carbonell i Cortés, Ovidi. 2014. *Los Enfoques Culturales en la Traducción Literaria*. Castellón: Universitat Jaume I.

Cercas, Javier. 2001. *Soldados de Salamina*. Barcelona: Tusquets Editores.

Cercas, Javier. 2002. *Soldati di Salamina*. Translation Pino Cacucci. Parma: Ugo Guanda editore.

Coseriu, Eugenio. 1981. *Lecciones de Lingüística General*. Madrid: Gredos.

Curell, Clara, Cristina G. de Uriarte and José M. Oliver. 2010. Viajes narrados y palabras viajeras: Voces españolas en los relatos de exploración franceses de principios del siglo XX. In Ana C. Santos ed. *Descontinuidades e Confluências de Olhares nos Estudos Francófonos.* Faro: Universidade do Algarve, 47–56.

Degerstedt, Andrea. 2013. *Guardar o no Guardar, una Cuestión de Prestar: Un Estudio de Neologismos y Préstamos y su Inclusión en el Diccionario de la Real Academia Española*. Uppsala: University of Uppsala.

Díaz Larios, Luís Felipe. 2007. Los libros de viajes y la traducción cultural. In Francisco Lafarga, Pedro S. Méndez Robles and Alfonso Saura Sánchez eds., 123–135.

Díaz Prieto, Petra. 1998. ¿Son los anglicismos el camino del spanglish? *Estudios Humanísticos Filología* 20: 163–177.

Even-Zohar, Itamar. 1990. Introduction [to Polysystem Studies]. Polysystem studies. *Poetics Today* 11/1: 1–6.

Gheno, Vera. 2019. The Italian-English "cocktail" on Italian Social Networks. *Quaderni di Linguistica e Studi Orientali* 5: 45–94.

Grochowska, Anna. 2010. La pastasciutta non è più trendy? Anglicismi di lusso nell'italiano contemporaneo. *ANNALES Universitatis Mariae Curie Sklodowska. Lublin* 2: 43–59.

Hatim, Basil and Ian Mason. 1995. *Teoría de la Traducción: Una Aproximación al Discurso*. Barcelona: Ariel.

Hervey, Sandorn and Ian Higgins. 1992. *Thinking Translation. A Course in Translation Method: French to English*. London: Routledge.

Lafarga, Francisco, Pedro S. Méndez Robles and Alfonso Saura Sánchez eds. *Literatura de Viajes y Traducción*. Granada: Editorial Comares.

Lope Blanch, Juan M. 1972. El concepto de prestigio y la norma lingüística del español. *Anuario de Letras. Lingüística y Filología* 10: 29–46.

Mangiron, Carme. 2006. *El Tractament dels Referents Culturals a les Traduccions de la Novel·la Botxan: La Interacció entre els Elements Textuals i Extratextuals*. Barcelona: Universitat Autònoma de Barcelona dissertation.

Mattioli, Virginia. 2018. *Los Extranjerismos como Referentes Culturales en la Literatura Traducida y la Literatura de Viajes: Propuesta Metodológica y Análisis Traductológico Basado en Corpus*. Castellón: Universitat Jaume I dissertation.

Molina Martínez, Lucía. 2006. *El Otoño del Pingüino*. Castellón: Publicacions de la Universitat Jaume I.

Newmark, Peter. 1988. *A Textbook of Translation*. New York: Prentice Hall.

Nida, Eugene. 1964. *Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating*. Leiden: Brill Archive.

Ortega Román, Juan José. 2006. La descripción en el relato de viajes: Los tópicos. *Revista de Filología Románica* 4: 207–232.

Osinaga, Itziar Enekoitz. 1999. *Carbonell i Cortés, Ovidi. Traducción y Cultura: De la Ideología al Texto*. Salamanca: Ediciones Colegio de España.

Pickford, Susan and Alison E. Martin. 2013. Introduction: Travel writing, translation and world literature. *InTRAlinea*. http://www.intralinea.org/specials/article/1963 (29 November, 2021.)

Piglia, Ricardo. 2010. *Blanco Nocturno*. Barcelona: Anagrama.
Piglia, Ricardo. 2011. *Bersaglio Notturno*. Translation Pino Caccuci. Milano: Feltrinelli.
Polezzi, Loredana. 2012. Translation and migration. *Translation Studies* 5/3: 345–356.
Porras Castro, Soledad. 1999. Sociolingüística del italiano contemporáneo: Hispanismos y americanismos. In Pedro Luis Ladrón de Guevara Mellado, Giuseppina Mascali and Pablo Zamora Muñoz eds. *Homenaje al Profesor Trigueros Cano.* Murcia: EDITUM, 603–618.
Real Academia Española. 2001. *Diccionario de la Lengua Española*. http://rae.es/drae/ (3 May, 2022.)
Slebus, Eva. 2012. *La Percepción y las Actitudes Lingüísticas hacia el Castellano Peninsular y distintas Variantes del Español Hispanoamericano.* Utrecht: The University of Utrecht dissertation.
Smecca, Paola. 2003. Cultural migrations in France and Italy: Travel literature from translation to genre. *Traduction, Terminologie et Redaction* 16/2: 45–72.
Trivedi, Harish. 2005. Translating culture vs. cultural translation. *91st Meridian* 4/1. http://iwp.uiowa.edu/91st/vol4-num1/translating-culture-vs-cultural-translation (29 November, 2021.)
Tymoczko, Maria. 2006. Reconceptualizing translation theory: Integrating non-Western thought about translation. *Translating Others* 1: 13–32.
Venuti, Lawrence. 1995. *The Translator's Invisibility*. London: Routledge.
Xianbin, He. 2007. Translation norms and the translator's agency. *SKASE Journal of Translation and Interpretation* 2/1: 24–29.

Corresponding author
Virginia Mattioli
Spain
E-mail: virginiamattioli@gmail.com

received: November 2021
accepted: June 2022
published online: July 2022

RiCL Research in Corpus Linguistics

Evaluating stance annotation of *Twitter* data

Vasiliki Simaki – Eleni Seitanidi – Carita Paradis
Lund University / Sweden

Abstract – Taking stance towards any topic, event or idea is a common phenomenon on *Twitter* and social media in general. *Twitter* users express their opinions about different matters and assess other people's opinions in various discursive ways. The identification and analysis of the linguistic ways that people use to take different stances leads to a better understanding of the language and user behaviour on *Twitter*. Stance is a multidimensional concept involving a broad range of related notions such as modality, evaluation and sentiment. In this study, we annotate data from *Twitter* using six notional stance categories —contrariety, hypotheticality, necessity, prediction, source of knowledge and uncertainty— following a comprehensive annotation protocol including inter-coder reliability measurements. The relatively low agreement between annotators highlighted the challenges that the task entailed, which made us question the inter-annotator agreement score as a reliable measurement of annotation quality of notional categories. The nature of the data, the difficulty of the stance annotation task and the type of stance categories are discussed, and potential solutions are suggested.

Keywords – stance-taking; social media discourse; corpus annotation; inter-coder reliability

1. INTRODUCTION[1]

The development of social media platforms has given rise to a new type of discourse serving different purposes. The platforms are used by different actors to express opinions and assess other people's opinions but also to construct and establish their online identity over time. Despite their similarities, each social media platform has a different character and slightly different policies. These conditions have repercussions on how the platform users communicate. Especially in the case of *Twitter*, users are restricted to a specific tweet size and specific interaction functions —reply, like, retweet and share— which naturally affect the nature of the discourse. Stance-taking in tweets is pervasive. Expressions of stance are used to promote, reinforce or mitigate the communicative goals of the users such as, for instance, to search for information or make information more visible (Zappavigna 2012: 50ff.). Conversational practices through the various *Twitter* features, such as retweet, reply and mentions, have emerged (Boyd *et al.* 2010), and as

[1] The authors would like to thank two anonymous reviewers for their comments. This research was supported by the *Kamprad Family Foundation* (Reference No: 20180178).

Honey and Herring (2009) point out, opinions, sentiments and stances are present in such interactions. The study of the discursive ways that *Twitter* users employ to communicate their stances offers important insights about users' behaviour and language use.

Stance and stance-taking are concepts strongly related to modality and sentiment that have been widely studied in different research fields and for various purposes. Kaltenböck *et al.* (2020: 1) define stance as

> the way in which speakers express points of view, attitudes, feelings and evaluations, and position themselves in relation to some proposition (i.e. subjectivity) and to other speech participants (i.e. intersubjectivity) and their particular stances.

A similar definition is provided for stance-taking in Simaki *et al.* (2020: 217) as

> the way speakers position themselves in relation to their own or other people's beliefs, opinions and statements about things or ideas in ongoing communicative interaction with other speakers.

Based on this definition, a stance framework with ten notional categories such as certainty, contrariety and necessity, among others, was introduced in Simaki *et al.* (2020). A general framework consisting of stance concepts that go beyond pro/con statements has the potential of important advances in stance studies in corpus pragmatics, computational linguistics, content analysis and other relevant disciplines. However, the annotation of texts using this stance framework is challenging since complexity, subjectivity and the background of the annotator can affect the annotation results and, consequently, the reliability of the dataset.

In this study, we test the validity of the abovementioned stance framework in order to show how suitable our categories are in a stance analysis task. Our stance framework was initially tested in data from blogs, and for this task we continued working with data from *Twitter*, as these data types fall within the social media discourse genre in the broad sense but are different in a range of ways from blog texts. Our purpose is to identify stance and attribute a stance label to the selected data, but we acknowledge the fact that this might not be possible for every tweet included in the data set. For this reason, in addition to the six stance categories that we used, namely, contrariety, hypotheticality, necessity, prediction, source of knowledge and uncertainty, we introduced a seventh category, entitled 'no label', which included those tweets that could not be attributed to any of the

other stance labels.[2] The *Twitter* data in the study were annotated by two experts (annotators A and B), and the inter-annotator agreement was calculated. The relatively low level of agreement between the annotators led us to a broader discussion of discourse annotation, inter-annotator agreement measurements and what is considered to be an acceptable agreement level that ensures the reliability of the annotated data. The particular aims of this study are:

1. to evaluate the stance framework on annotated *Twitter* data of a wide thematic range;
2. to identify patterns and/or possible problems of the annotation scheme;
3. to propose solutions to improve the annotation results in the future;
4. to describe and analyse the complexity and the different components of tweets annotated as 'no label' for the refinement and improvement of the stance framework and annotation protocol.

2. Background work

As a result of the expansion of social media platforms, social media discourse has become the focus of research from various perspectives in linguistics and other disciplines. The analysis of this discourse type can be a challenging task, because of ethical, formatting and language issues that may arise (Hernández 2014), as well issues related to the authors' identity and communicative purposes (Yus 2011, 2016). *Twitter* data is special in many ways regarding the relations among users and the features that are available. *Twitter* users establish social relationships based on the notion of 'following' (the user has followers and follows other accounts), and this affects the tweets that are shown in their timeline, which has an impact on their network. When it comes to the platform's features, the @ symbol is used for addressivity/communicative purposes among users and the # symbol as a feature of searchable tweets/conversations. These and other *Twitter* features have been extensively studied, especially hashtags (sequences starting with the # symbol) and their function that enables users to search for specific content and make comments searchable for others (Zappavigna 2015; Zhu 2016). A new type of publicness has emerged from *Twitter* with users presenting information of personal relevance (Schmidt 2014). *Twitter* is also used to create communities and networks sharing common

[2] See Section 3 for a detailed description and examples of all categories and labels.

experiences and/or similar values, and in such environments stance-taking is pervasive (Zappavigna and Martin 2018).

Broadly speaking, stance-taking is the way people use language to position themselves, express their opinions and assess their own and other people's messages (Du Bois 2007). It has been studied in various contexts, and a whole range of aspects are involved, these including modality (Facchinetti *et al.* 2003; Marín-Arrese *et al.* 2014), evaluation (Hidalgo-Downing 2012; Fuoli 2018), evidentiality (Ekberg and Paradis 2009), subjectivity/intersubjectivity (Verhagen 2005; Marín-Arrese 2017) and sentiment (Taboada 2016). The analysis of speaker stance is a vibrant area in language sciences, with many studies aiming to understand better its role in human communication (Hunston and Thompson 2000; Berman *et al.* 2002) and its association to social roles, identities, interpersonal and social relationships (Jaffe 2009), while others focus on stance phenomena in specific types of discourse (Hyland 2005; Biber 2006; Perrin 2012), including social media discourse (Jacknick and Avni 2017), specific stance-taking expressions (Paradis 2003) and discourse markers that are strongly related to stance (Traugott 2020). Apart from the qualitative approaches, corpus-based methodologies offer important insights into the identification of stance and stance expressions in discourse. Such methods and tools offer the possibility to investigate stance-taking in large amounts of data and perform statistical tasks and analyses to identify patterns in the data across time and discourse types (Alonso Ameida 2015). Stance has also been studied from a computational perspective (Ghosh *et al*. 2019; Küçük and Can 2020) with many researchers addressing stance as a binary phenomenon of the speaker's pro/con positioning in relation to a topic, an idea or an event (AlDayel and Magdy 2021). Stance annotation, in particular, has also been studied extensively with researchers aiming at creating as comprehensive annotation systems and tools as possible, which allows to use the annotation for automatic stance detection and classification (Kucher *et al.* 2016). Such tasks are performed in data extracted from ideological forum debates (Hasan and Ng 2014), news articles (Ferreira and Vlachos 2016), academic text data (Faulkner 2014) or other social media sources (Mohammad *et al.* 2016; Pamungkas *et al.* 2019).

In Simaki *et al.* (2020), the point of departure is a notional definition of speaker stance rather than a lexical one. According to this definition, discussed in Section 1 above, the concept of stance is defined as a psychological state involving speakers' beliefs and attitudes, stance-taking as human performance in communication and expressions of

stance as the constructions used for stance-taking in discourse. As a result, and based on the literature in the field, an original stance framework consisting of ten notional stance categories was proposed.[3] These categories were manually identified and attributed to utterances extracted from blogs thematically related to the 2016 UK referendum. The final output of this procedure resulted in the *Brexit Blog Corpus* (BBC).[4] Simaki *et al.* (2020) showed that stance-taking is common practice in discussions of controversial political matters such as the Brexit. The distribution of the categories showed that contrariety was the most frequent category in the corpus, while the category of volition was the least frequent one. The presence of more than one instance of stance-taking in the same utterance was also shown to be a frequent phenomenon. The calculation of the inter-coder reliability showed good agreement scores for the categories of contrariety, hypotheticality, necessity and uncertainty.

In subsequent studies, the BBC was computationally (Simaki *et al.* 2017a) and statistically (Simaki *et al.* 2018a) evaluated in order to test the framework's efficacy and to provide new insights about linguistic patterns for the identification of stance in discourse in future work. In Simaki *et al.* (2019), the aim was to identify specific constructions that are related to the six most frequent stances in the BBC categories. A quantitative analysis of the annotated corpus data and a meta-annotation procedure to identify lexical forms (stance markers) that are stance-specific for each category were performed. The results of the two techniques were then compared, and a list of constructions of stance-related discourse as particularly salient expressions of each stance type was proposed.[5] Part of this list is used in the present study, as will be shown in Section 4.

3. Methodology

In this study, our hypothesis was that the stance framework mentioned above is suitable for the analysis of stance in discourse and its use can be generalised to social media discourse types other than blogs and a wide variety of topics. For that purpose, we used texts retrieved from *Twitter* which were extracted on the basis of specific criteria from a social media corpus (see Section 4). We selected *Twitter* as the source of data for our

[3] See the full framework with a brief description and examples for each category in Appendix 1.
[4] https://snd.gu.se/en/catalogue/study/snd1037
[5] These constructions are presented in Appendix 2.

study since tweets vary from blog texts, the most important difference being the character limitation of the tweet in contrast to blog texts, which can be as long as the blog author wants. In addition, *Twitter* is frequently the source of data in which researchers from different disciplines dive into to explore people's ideas, beliefs and opinions about various topics, and we have prior experience with the particularities and challenges of such data type.

We used the six most frequent stance categories distinguished in the stance framework (Simaki *et al.* 2020), namely, contrariety, hypotheticality, necessity, prediction, source of knowledge and uncertainty. The category of contrariety includes instances where the authors express a compromising/contrastive opinion (e.g., *Hate the end result, but #thegame always delivers. always. best rivalry in sports*).[6] Hypotheticality is attested in utterances where authors express a possible consequence of a condition, mostly formulated with conditional clauses (e.g., *If you use this Kim Kardashian hashtag thing it's an instant unfollow*). Necessity includes cases in which authors express requests, recommendations, instructions or obligations (e.g., *I really need to start utilizing a day minder*). Prediction is attested when authors make a guess or a conjecture about a future event (e.g., *@lazycat99 I knew someone would catch that reference. Well done!*). Source of knowledge occurs when authors express the origin of what they say (e.g., *One mustn't be much concerned with living, but with living well... Socrates to Crito, in Plato, 'Crito', 48b*). Finally, the category of uncertainty concerns authors' doubt regarding the likelihood of what they say (e.g., *Stand up special starts in 20 mins. I think 7 on West Coast. 10 on East. I actually have no fucking idea*). As already stated (see Section 1), in the present study, we introduced another category, 'no label', which deals with tweets that did not fit in any of the abovementioned categories. This includes neutral statements (e.g., *rt @ankhmarketing: ms. lauryn hill [@mslaurynhill] performing live may 12th [@thewarfield!!] @goldenvoicesf*), questions (e.g., *@amwalkush @britenyc but we are still going to decorate gourds, right?*), ambiguous and/or illegible tweets (e.g., *Me. Stretch. Hollywood. rt @will_blackmon: I wear a 3 piece suit in a cab son. Who needs a limo!*) and tweets expressing sentiment (e.g., *So grateful for u all and ur kind comments. May this brighten ur day. Love u! #standbyyou*) or more than one stance (e.g., *@jjenas8 You might be right but you're wrong*). Two annotators annotated the data (see Section 5.1), and the inter-annotator agreement was calculated.

[6] Unless otherwise stated all instances have been retrieved from dataset used in this study.

4. CORPUS DESCRIPTION

We use data from the *Twitter* part of the social media corpus used in Simaki *et al*. (2017b). Simaki *et al.*'s (2017b) corpus was compiled with data from the official *Facebook* and *Twitter* profiles of public figures such as actors, authors or athletes. It was manually annotated with the authors' sociodemographic information such as their gender, age, profession and any other additional information available as, for instance, their educational background. In contrast to the BBC, this corpus consists of texts on various topics, such as personal branding, social and political matters, nature, etc. The corpus was compiled from September to December 2015 at the same time the BBC was build. It includes texts from 838 different authors (535 male and 303 female authors) and its overall size is of 13.4 million words distributed in 721,033 entries. The data were further processed and normalised and, as a result, features typical of *Twitter* discourse (e.g., multimodality, the use of upper/lower case letters, hashtags or emojis/emoticons, among others) were excluded and, therefore, are also disregarded in the present study. However, some features such as hashtags (#), mentions (@) and links have been included in the data.

In Simaki *et al.* (2019), a list of stance markers for each stance category was compiled containing both stance-related forms, such as *but, if, must,* and forms that do not unambiguously evoke a specific type of stance but are stance-related in the sense that they occur frequently in long sequences that express stance, such as *then* (e.g., *If you're not willing to risk it all then you do not want it bad enough*) and *would* (e.g., *It would be cute if they didn't draw on me*) that are frequent forms in hypothetical sentences. The markers are based on a two-fold analysis of the BBC data: first, the extraction of the statistically significant lexical items per category and, second, the identification of the stance-related lexical chunks by one of the annotators, who —five months before this task— had conducted the initial annotation task. The findings from both analyses were combined and the results are shown in Appendix 2. For the present study, we refined that list by excluding forms that would create noise in the data selection process such as *I, be, is, have* and *it*. To avoid a high number of neutral or irrelevant tweets, we selected texts from the *Twitter* set of the corpus in which at least one stance marker from the refined list was present. This list contains 20 markers, and 1,000 tweets were extracted. This was possible for many of the markers that are used frequently in tweets but, in some cases, the search retrieved fewer tweets. In Table 1, we present the list of the stance markers that

were searched for in the data, the corresponding stance categories for these markers and the number of tweets extracted per marker.

Stance category	Stance marker	Number of tweets
Contrariety	*But*	1,000
	Than	1,000
	While	436
Hypotheticality	*Could*	1,000
	If	1,000
	Would	1,000
	Then	819
Necessity	*Need*	1,000
	Must	396
	Needs	259
	Should	1,000
Prediction	*Will*	1,000
Source of knowledge	*As*	1,000
	Said	582
	Show	1,000
	That	1,000
Uncertainty	*Think*	1,000
	Might	350
	Maybe	337
	Probably	168
Uncertainty/ prediction	*May*	521
	Total	**15,868**

Table 1: Stance markers used for the extraction of the data listed according to the stance categories they pertain to, and the number of the tweets extracted

As illustrated in Table 1, the total size of the dataset is 15,868 texts (274,697 words). The markers *may*, *maybe*, *might*, *must*, *needs*, *probably*, *said*, *then* and *while* were limited in number (fewer than 1,000) and, thus, all tweets in which they were present were extracted. The relevance of these stance markers to the annotation results and our research findings will be discussed in Section 6.

5. Corpus annotation and results

In this section, we describe the annotation procedure and the annotation results from the pilot and the final annotation rounds.

5.1. Corpus annotation procedure

The annotation of the data was carried out by two annotators with a background in linguistics. More specifically, annotator A holds a PhD in linguistics and computational linguistics, whereas annotator B holds a Master in English applied linguistics. A comprehensive annotation protocol was introduced, comprising six steps, as presented in Table 2.

1	Presentation of the main concept, the stance categories and familiarisation with previous studies.
2	New label for current task: 'no label' category in which neutral statements, questions, ambiguous and/or illegible tweets and tweets with sentiment or more than one stance are stored.
3	Discussion between Annotator A (research expert) and Annotator B (research assistant) about the task.
4	Pilot annotation of 664 tweets by annotator A and Annotator B.
5	Discussion between Annotators A and B about conflicting assessments/ambiguous cases.
6	Annotation of 6,659 tweets by Annotator A and 15,868 by Annotator B.

Table 2: The annotation protocol followed in the study

As shown in Table 2, firstly, annotator A explained the main concept, the stance framework and the categories. Instructions about the annotation process were also provided, so that both annotators would base their decisions on the overall meaning of each text and would not merely rely on the potential presence of a specific stance marker. Annotator B studied previous work to become familiar with the task. Secondly, a new category was added: the 'no label' category. Annotators attributed this label to neutral statements that do not express any stance, ambiguous or illegible tweets, tweets with more than one stance and tweets expressing sentiment but not stance. In addition, we did not exclude the tweets in which a question mark was present, as in many cases it is not used to form a question (e.g., *@imsoforserious calling someone ugly, stupid or a cunt is hardly criticism. but thanks for trying to teach me something super obvious (?)*). The idea to add this category stems from Simaki *et al.* (2018b), in which many texts were stance-free, neutral, expressing sentiment or irrelevant. Further analysis of this category will provide feedback about the discourse of *Twitter* and improve the stance annotation process. In the third step of the annotation process, annotators A and B discussed the task while the fourth step was a pilot annotation round of 664 tweets by both annotators. In step five, after the pilot round, the two annotators discussed the challenges of the task, problematised conflicting or ambiguous tweets, and problems were resolved. Finally, in step 6, the final annotation round was conducted.

5.2. Pilot round annotation results

The reliability of the annotated set of 664 tweets in the pilot round was tested by calculating the level of agreement between the annotations by annotators A and B. We used the coefficient kappa (Cohen 1960) to calculate the inter-annotator agreement score with a confidence level 95 per cent. The results are shown in Table 3 which provides the distribution of the annotated tweets in each stance category, and the inter-annotator agreement score. The highest level of agreement between the two annotators was achieved for the source of knowledge category (0.77), followed by the necessity (0.59) and contrariety (0.58) categories. The overall inter-annotator agreement for this set of tweets is 0.54, which can be characterised as moderate according to Landis and Koch (1977).

		Annotator A								
	Categories	**Contrariety**	**Hypotheticality**	**Necessity**	**Prediction**	**Source of knowledge**	**Uncertainty**	**No label**	**Total**	**Kappa**
Annotator B	**Contrariety**	71	5	1	6	8	7	21	119	**0.58**
	Hypotheticality	4	49	1	3	2	2	7	68	0.53
	Necessity	7	38	91	9	2	12	15	174	**0.59**
	Prediction	3	0	0	31	1	15	4	54	0.48
	Source of knowledge	7	2	5	6	86	0	8	114	**0.77**
	Uncertainty	4	2	1	8	1	63	34	113	0.51
	No label	4	4	1	1	0	1	11	22	0.13
	Total:	100	100	100	64	100	100	100	**664**	**0.54**

Table 3: Annotation results of the pilot round and kappa scores

5.3. Final round annotation results

After steps four and five, the final annotation round was carried out: 6,659 tweets were annotated by annotator A and 15,868 tweets by annotator B. Table 4 shows the overall results of the annotation. As can be noticed, the ‘no label’ category is the largest category according to the annotations of both annotators. This category is more than twice as large when compared to the rest of the annotated categories, which shows the extent to which our stance annotation criteria did not apply. For annotator A, the most frequent categories are uncertainty, contrariety and necessity. For annotator B, necessity is the most frequent

type of stance, which is followed by contrariety and hypotheticality. For both annotators, prediction is the least frequent category.

Stance categories	Annotator A	Annotator B
Contrariety	687	1,632
Hypotheticality	386	1,207
Necessity	628	2,235
Prediction	64	278
Source of knowledge	134	809
Uncertainty	730	1,034
No label	4,030	8,673
Total	**6,659**	**15,868**

Table 4: Final annotation round results

We, then, calculated the interrater reliability of the annotated tweets, which is shown in Table 5.

		Annotator A								
	Categories	**Contrariety**	**Hypotheticality**	**Necessity**	**Prediction**	**Source of knowledge**	**Uncertainty**	**No label**	**Total**	**kappa**
Annotator B	**Contrariety**	439	3	9	5	2	12	332	802	**0.54**
	Hypotheticality	5	164	2	3	0	6	224	404	0.38
	Necessity	5	66	383	2	0	15	397	868	0.45
	Prediction	0	0	3	17	1	12	69	102	0.19
	Source of knowledge	15	11	20	4	91	33	202	376	0.34
	Uncertainty	4	13	15	12	0	429	187	660	**0.57**
	No label	219	129	196	21	40	223	2,619	3,447	0.32
	Total:	687	386	628	64	134	730	4,030	**6,659**	**0.42**

Table 5: Final annotation results and kappa scores

As can be noticed, the kappa score for the total set of annotated data is 0.42, which is a much lower score than the score in the pilot round. In this set, we achieved the highest inter-annotator agreement score for the uncertainty category (0.57), as well as the second highest score for the contrariety category (0.54). Among the labelled tweets, these two categories are the most frequent ones. Interestingly, source of knowledge, which was the stance category with the highest inter-annotator agreement score in the pilot round, shows a lower kappa score in this round (0.34). The most important finding is the frequency of the 'no label' category in this annotation round (55% of the corpus). However, the low agreement score (0.32) on the tweets grouped in this category suggests that the annotators faced difficulties in applying the annotation instructions in the same way. This difficulty can be due to the relatively high level of the subjectivity of the task since the categories

are notional rather than being identified through lexical items. In comparison with the very low kappa score in the pilot round (0.13), in this round, the kappa indicates a better agreement score (0.32).

The low overall kappa score in the final annotation round led to the implementation of alternative measures which have more advantages regarding the type of data that they support and the handling of the missing data. We calculated the Krippendorff's Alpha coefficient (K alpha; Krippendorff 2011), which is another standard and relevant metric for the calculation of the interrater reliability as a reference metric. In addition, we calculated the Gwet's AC_1 coefficient (Gwet 2002), a more recent metric which has been suggested as a more robust solution to evaluate annotations of discourse data, in which skewed data and variability in the distribution of categories are quite common phenomena (Hoek and Scholman 2017). Table 6 shows, the inter-annotator agreement scores, which are calculated by using three different metrics. The results show that the kappa and K alpha scores have similar values (0.42 and 0.41, respectively), while Gwet's AC_1 shows a higher score (0.58). These results gave rise to methodological considerations regarding the annotation of discourse data and the annotated data reliability and quality. This will be discussed in Section 7.

Metric	**Score**
Kappa	0.42
K alpha	0.41
Gwet's AC_1	0.58

Table 6: Results of the different metrics

6. Analysis of the annotation results

When it comes to the analysis of the annotated data, we start with the frequency of the six stance categories. As shown in Table 5, the most frequent stances for annotator A were uncertainty, contrariety, necessity and hypotheticality. For annotator B, necessity was the most frequent stance, followed by contrariety, hypotheticality and uncertainty. Examples (1)–(4) illustrate the most frequent categories in the annotated data.

(1) (@carlykimmel I think we get 5 years of this. It ends in 7th grade, just like my grandmother promised me. (Uncertainty)

(2) @chelseaolson3 @andygrammer I did find this but haven't used it yet. (Contrariety)

(3) I must apologise straight away for leaving the question mark off the end of my previous tweet. (Necessity)

(4) Damn it! If I go one week without seeing game of thrones I have to start from the beginning again. (Hypotheticality)

These tweets are examples in which both annotators agreed on their label attribution. In these categories, the best inter-annotator agreement score was achieved (see Table 5). The high number of tweets annotated as 'no label' is of great interest as well. It turns out that this was the largest category of the dataset with more than half of the data annotated as 'no label'. The annotators faced several challenges during the annotation of the data that led them to attribute this label to different reasons: the presence of symbols and/or special characters that made tweets difficult to comprehend, tweets consisting only of a hyperlink, incomprehensible abbreviations, slang language, the absence of enough context and the absence of any stance category in many cases. Additionally, the selected texts cover a wide thematic range, where stance-taking may not always be among the main communicative purpose of the tweeter. This contrasts with the BBC, where the discussion about a controversial political matter invites people to express their stance in a bolder manner. Therefore, identifying stance in *Twitter* data creates more noise in our corpus, with content that cannot be grouped under the predefined stance categories.

A closer look at the 'no label' data confirms the diversity of this category and various patterns may be observed. According to our guidelines, tweets that express neutral and stance-free statements should be in this category. This type of tweets is frequent, and two examples of neutral and stance-free tweets are shown in (5)–(6).

(5) The economy added 280,000 jobs in May marking 63 consecutive months of private-sector job growth.

(6) FYI I just sat down to google "how to use pomade" and somehow tweeted that.

In these examples, tweeters either make a neutral statement or describe aspects of their lives without taking any stance. More specifically, the authors share neutral information probably derived from news sources (cf. 5) or describe their everyday experiences (cf. 6). In many cases, the tweets are narrated in a confessional/emotional tone to forge connection with their followers, bond with them and/or increase their network. Tweets in

which gratitude, love and wishes are expressed by public figures to their followers were quite frequent in our data. Some examples are provided in (7)–(9).

(7) That was fun times #moa!!!! love you mean it Minneapolis!!! @mallofamerica

(8) That plane saga made my night. Happy thanksgiving to you & yours! rt @briankoppelman happy tg. have u been following @theyearofelan tonight?

(9) Happy Valentines Day! May you love, be loved and make love, all in excess!

These examples illustrate tweets expressing sentiments, such as gratitude, love, appreciation and enthusiasm that public figures express to their fans. This type of tweets usually creates interaction and followers respond with likes, retweets and replies. The initial tweet becomes more visible, while the author builds stronger ties with their followers and gets more followers. The follower, in turn, gets the chance to establish a 'real' connection with the public figure they admire. As a result, the public figure has a larger and more loyal audience to which self-branding and promoting strategies can be efficient, as shown in (10)–(12).

(10) We are in Orlando, fl @waltdisneyworld for an amazing event. social media moms celebration. I must have spoken well last year. I'm back again

(11) Looking for some new music for the weekend? check my #liveinthefuture top 10 chart at beatport *HYPERLINK*

(12) Seriously the best cafe in California is @cafegratitudevb [...] if you haven't already tried it you need?? *HYPERLINK*

Examples serving the purpose mentioned above can also be attested and are annotated with a stance label, but most tweets related to promoting and self-branding content or providing advice about health and lifestyle choices were grouped in the 'no label' category, even in cases where indications of stance could be detected, as in (13) and (14).

(13) Free tickets, a free round-trip flight and free swag? What else could a steelers fan ask for?! Enter here!

(14) Be strong & courageous. do not be terrified or discouraged, for the lord your god will be with you wherever you go. -josh 1:9 be #blessed !!

In (13), the public figure urges their followers to join a competition to win free tickets, while in (14) lifestyle/religious advice is given. In both examples, the authors recommend their followers about specific choices, and the necessity label could be used in both tweets, but due to their overall meaning, the utterances, do not only express necessity. More specifically, in (13), the text includes two questions, with the second question also expressing uncertainty, and it ends up with an exhortation to the followers to join the competition. In (14), adding to the recommendation expressed in the imperative, we can also identify prediction (*…the lord will be with you…*) and source of knowledge (*-josh 1:9*). The co-occurrence of different types of stances in the same text was already observed in the analysis of the BBC (Simaki *et al.* 2020) and is also a frequent phenomenon in the present data. In (15)–(17), this co-occurring pattern may be observed.

(15) **Need** to sleep **but** my stupid brain won't shut offffff. Hummmblfukkdstfjff *HYPERLINK*. (Necessity and contrariety)

(16) #morningjah "you change **if** you change from babylon to rasta, **but** you can't change from rasta to anything". (cont) *HYPERLINK*. (Hypotheticality and contrariety)

(17) **Don't know if** you guys saw @hitrecordjoe be one of the first to do it **but** he did @nickiminaj like nobody's biz. go!x *HYPERLINK*. (Uncertainty, hypotheticality and contrariety)

Other patterns of tweets characterised as 'no label' are texts with stance-taking but, since the text (or part of it) is a question, they have been excluded. Some examples are provided in (18)–(20).

(18) I'm sorry but did you see my last post? My fans care about others in a manner I can't even begin to explain. proud. Let's change the world!

(19) Heartbreaker but the entire group is still alive. why not us? #ibelieve

(20) If obama's asia trip wasn't suspicious, why are all of his meetings taking place while Americans are asleep? *HYPERLINK*

Finally, new constructions related to stance were attested in this category. We identified various patterns of commonly used expressions that can be associated to our stance categories or form new ones. For instance, an interesting pattern is the 'not sure' construction that can be aligned with the uncertainty category, but it frequently co-occurs

with other stances or is part of a question. While it certainly evokes a sense of uncertainty to the whole text, this pattern made the annotators doubt as regards the strength of the 'not sure' construction in dominating the overall meaning of the text, especially when other stances could also be identified. As a result, such cases were annotated as 'no label'. Some examples of this pattern are provided in (21)–(23).

(21) @jh0ps maybe...maybe not....probably maybe tho...but, maybe not also...dunno...could have...not sure...:):) (say hi next time!)

(22) @andavis1 college Wasn't right for me. Not sure what you mean about venture capital but a Boston based firm, spark, invested in jelly.

(23) I do. Not sure if I'm allowed to tell my prediction. I will check with NBC. Back later. RT @lolabeauty33 Any favorite acts from yesterday???

In (21), the tweet is illegible and not clear, while the contrariety marker *but* is present. In (22)–(23) other stances and sentiments may also be identified. Constructions, such as the 'not sure' construction, are strongly related to stance and should be studied in depth as they can enrich not only our stance markers list, but also our stance categories. For instance, and in contrast to the 'not sure' construction, the certainty category, for which we identified examples of the 'I'm sure' construction in the data, could be added. Examples are provided in (24)–(25).

(24) @1nataliemaines: I'm sure this haircut will be coming back around any day now. I think you should have it now.

(25) nicert @stephpalmer15: @mooremaya are you planning on coming to #passion2013? I'm sure @lecrae would share the stage!

Nevertheless, we also need to address the strength of the certainty that the 'I'm sure' construction carries in connection to the occurrence of other stances: prediction and necessity in (24) or the presence of the question in (25). In the present study, we have not addressed the issue of the presence of expressions of different stances in the same text, as we focus on the annotation process, the observation and the analysis of the results.

Overall, the annotation results mostly confirm the validity of our stance framework: the six stance categories tested here are attested in the *Twitter* data and, especially for the cases of uncertainty and contrariety, we observed a moderate but acceptable level of inter-annotator agreement. The overall agreement score (0.42) highlights the challenges of the

stance annotation of the *Twitter* data, but also the potential of such a task. Sentiments are also present in the data, and constructions related to sentiment can be identified for a more in-depth linguistic analysis of the data.

In addition to the analysis of the annotation results, we studied whether the stance label which was attributed by the annotators to each text corresponded to the stance that, according to Table 1, the marker that was present in this text indicated. The goal of this task was first to test whether the presence of each of the stance markers provides a robust indication for the overall meaning (stance-taking or not) of the text and, second, to confirm whether each selected marker was perceived as related to a specific stance category by the annotators. Our hypothesis has been that the frequency of the stance types in the data does not only reflect the way that *Twitter* authors position themselves in their text, but it is also linked to the selection of the data, which is based on the list of predefined stance markers described in Section 4. This list includes a range of stance markers from relatively clear ones (*but* and *if*) to items that do not refer to a specific stance category (*that*, *show* and *think*). We compared the annotation labels that the annotators attributed to the stance marker according to which text was selected to be part of the dataset. We also investigated whether the stance category related to each of the 21 stance markers coincided with the annotation label that the annotators decided to attribute to the tweet: for instance, are texts in which *as* and *but* (markers for source of knowledge and contrariety respectively) annotated as source of knowledge and contrariety? We then associated the annotations to the stance markers and the results are shown in Figure 1 below.

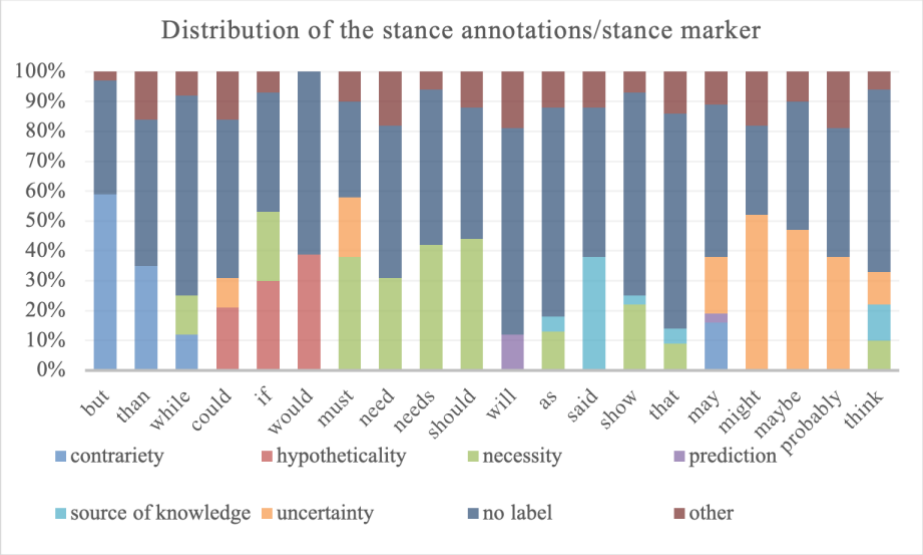

Figure 1: The percentage distribution of annotations per texts in which the same stance marker is present

Figure 1 shows how, in terms of frequency, the six stance categories and the 'no label' category are attributed to the data. For each marker subset, we assumed that the distribution of the labels would reflect the correlation of the annotations to the corresponding stance marker which, in turn, is associated with a given stance category. For instance, in the case of *but* (that may be a marker of contrariety), almost 60 per cent of the extracted data, in which *but* was present, was annotated as contrariety. The results here show that, to an extent, the well-established stance markers (*but, if* type) were annotated with the stance category to which they are strongly related (*but* with contrariety and *if* with hypotheticality), despite the high percentages of the 'no label' category in all subsets. Words that are less evident as markers of a specific stance, such as *as, that, think* and *will* are more rarely annotated with the anticipated stance label. For instance, *as* was identified as a marker for source of knowledge in Simaki *et al.* (2019), but, in the present study, texts containing *as* are mostly annotated (70%) as 'no label', and also as necessity (13%). Only 5 per cent of these texts were annotated as source of knowledge. This shows that the same marker is used to express different stance. Another example of multifunctionality and maybe text type/genre sensitivity is *show*, which in the BBC was used in source of knowledge constructions as a verb (e.g., *The data from the study show that…*). In the present data, it mostly refers to artistic performances and, despite the large number of 'no label' cases, it is frequently attested in necessity texts, where public figures encourage/urge their followers (e.g., *Dear everyone in Perth, u must see this show. It won best cabaret last year & it opens tomorrow for one week only!*). This phenomenon is due to the different types of content, topics and text types of both corpora, so we can assume that the same form that is identified as stance marker in one dataset does not work in the same way in a different dataset. Overall, the results in Figure 1 provide interesting insights about the validity of these stance markers when tested in a different dataset.

7. Stance annotation methodological considerations

Most of the discussion since the first annotation round of the BBC in Simaki *et al.* (2020) has been about the challenging nature of the stance annotation task. Difficulties were inevitable due to most of the framework's categories and the nature of the BBC (limited size and duplicates due to stance co-occurrences). Likewise, all efforts of quantitative and computational tasks within the given setting have possibly resulted in overfitting stance-related markers as only the BBC was tested. Thus, it becomes even more challenging to

evaluate the findings from those studies in a different setting, and the results provided in Table 5 confirm such a challenge.

Similarly, less encouraging results can still be discussed, and important insights with a potential to address them in future studies can be achieved. Methodological questions can be raised and method-related issues problematised as to the effectiveness of our annotation protocol in applying a notional stance scheme. In this study, the weak inter-annotator agreement made authors reflect on the annotation results, and more specifically on the metric used (Cohen's kappa) and how suitable this metric is for the task. For this purpose, after an extensive bibliographic search, we have concluded that there is no consensus on which metric is the most appropriate one for calculation of inter-coder reliability, despite all efforts to develop reliable metrics and tools. There are several recommendations for Cohen's kappa, which is among the most frequently used metric. A common issue that arises when calculating the reliability of annotated data in a scheme with more than two labels is that infrequent categories emerge from the annotation process, which leads to an uneven distribution of categories that produces unbalanced datasets, and subsequently leads to a lower reliability score (McHugh 2012). This is a common phenomenon in discourse annotation studies where similar distributions of categories between different types of discourse are not always feasible (Hoek and Scholman 2017).

The frequency of a specific label is due to the frequency of the type of relation it refers to (e.g., condition, reason, opposition, etc.). Discourse is also characterised by an uneven distribution of connective constructions that mark the various relations and link the different parts of a sentence. Some of these connectors are very frequent while others are less frequent, and the distribution of relation types that specific connectives mark may also vary. As Hoek and Scholman (2017: 1) state,

> annotators tend to agree more when annotating explicit coherence relations, which are signalled by a connector or cue phrase (*because*, *for this reason*) than when annotating implicit coherence relations, which contain no or less linguistic markers on which annotators can base their decision.

This is important, as such markers not only are explicitly mentioned, but they are also less prone to ambiguity, so they cannot easily be interpreted in an ambiguous way. In the present study, this has been confirmed in the three most frequent categories in the annotated data (contrariety, necessity and uncertainty). Markers that signal the

corresponding stance type, such as *but*, *need*, *must* and *might*, occur frequently. The annotators could identify and label those markers that were explicitly associated to these stances and, as a result, the highest degree of agreement scores was attested. In some of the other categories, we can argue that due to the lower prevalence of stance-related items, there is insufficient information in the data, not only for the annotators to make decisions in a structured and homogeneous way based on discriminate factors, but also for us to assess the annotators' ability. As a consequence, kappa may underestimate the true agreement (Hripcsak and Heitjan 2002). An interesting case is the hypotheticality category which, despite the highly discriminative item *if*, is about half as frequent as the contrariety or the uncertainty categories and shows a lower level of agreement (0.35).

These issues can influence a reliability measurement such as a kappa score, which seems to be very sensitive to typical characteristics of discourse data, such as the ones mentioned above. In those cases, the kappa paradox is attested (Feinstein and Cicchetti 1990); in other words, the values are sometimes relatively low, despite the high percentage of observed agreement. We considered the agreement percentage as an alternative measurement that is easy to calculate and interpret but, as Lombard *et al.* (2002) argue, it fails to account for agreement that occurs by chance. Instead, as shown in Table 6 (see Section 5.3), we used two other metrics to calculate the interrater reliability: 1) the K alpha (Krippendorff 2011), which is also a standard metric, and 2) the relatively new AC_1 measure (Gwet 2002), which is used to solve some of the problems in the Cohen's kappa. This metric estimates the agreements between annotators as they are not partly due to chance (expected agreement) and it is less affected by the prevalence of categories and the marginal probability than the Cohen's kappa. AC_1 shows a higher score (see Table 6), which is encouraging for future research and suggests that it can be an important alternative measure when it comes to the calculation of the interrater reliability of discourse data.

The calculations of the inter-annotator agreement raised another important question about the interpretation of the results: What can be considered as an acceptable level of reliability? Does 0.58 here indicate that our set of annotated data is a reliable value for replication and usability purposes? The answer to these questions is problematic and, as Neuendorf (2017: 168) summarises, there are no established standards and "coefficients that account for chance (e.g., Cohen's kappa) of .80 or greater would be acceptable to all, .60 or greater would be acceptable in most situations and, below that, there exists

disagreement". Landis and Koch (1977) suggest a scale for the interpretation of kappa scores that was originally designed for the medical field: 0.41–0.60 values signal a moderate agreement, 0.61–0.80 substantial agreement and 0.81–1 perfect agreement. Poesio (2004) suggests 0.80 as a threshold that ensures an annotation of reasonable quality. McHugh (2012) states that kappa is not very well supported for factors such as rater independence, which lowers the estimate of agreement excessively. In addition, the fact that this metric cannot be directly interpreted leads researchers to accept lower kappa values. When it comes to the publication of annotated corpora, Artstein and Poesio (2008) argue that setting a specific agreement threshold should not be a prerequisite as long as a detailed report on data collection methodology, statistical significance of agreement and agreement table are included in the data description. Our opinion, which is based on the experience with different discourse annotation tasks, is in line with Artstein's and Poesio's (2008). We agree that interrater reliability is an important indication of the quality of annotated data and that it is important to use such measurements, but it is always worthwhile taking the analysis a step beyond the interpretation of the agreement score and, in doing so, draw more insightful conclusions.

8. Conclusion

In this study, our goal has been to evaluate 1) Simaki *et al.*'s (2020) stance framework and 2) the suitability of our categories in a different social media text type. We selected *Twitter* texts in which at least one stance marker from a predefined list was present. The data were annotated by two annotators and the inter-annotator agreement score was calculated. The findings show that taking stance differs across different social media platforms and that the stance categories, which appear to be salient in blogs, are less salient in *Twitter*. Our prior experience with stance annotation showed that it is not possible to identify stance in every text since people use *Twitter* (and social media in general) for different communicative purposes. Thus, many tweets can be stance-free in the sense of expressing sentiments or asking for information, while other issues, such as ambiguity or just illegible content, may still be present. For this reason, we created an additional category to cater for all tweets that did not conform to any of the six stance categories. According to the annotation results, this 'no label' category emerged as the largest one, which made us question the suitability of the stance categories for *Twitter* data. However, a closer look to the data made us realise that it provides an excellent

benchmark to further explore and develop this framework. A refined annotation protocol, a more cautious filtering of the data and adjustments in the existing framework are likely to lead to more efficient annotation and more reliable data. An alternative approach can also be considered, namely, to address the disagreements in the annotations and resolve the conflicting cases. Overall, our study confirms that the stance categories analysed here can be identified and attributed in *Twitter* data, despite the challenges of the nature of the task. In a follow-up study, sentiments, functions such as self-branding and new stances or other phenomena can be incorporated to the framework. In addition, the emerging patterns in the 'no label' category can be further analysed, and new categories can be considered to enrich the existing ones. In this study, both annotators devoted a large amount of time to a laborious cognitive task. Especially relevant has been the annotators' fatigue due to the manual task and its possible effects on the annotations, dealing with the very complex concept that stance is. As for the quantitative analysis, the statistical outcome has not been as rewarding as we had hoped. Should these results determine the reliability of the task, or is there room to derive important insights about stance-taking on *Twitter* data? The relatively low interrater agreement highlighted challenges related to the nature of the task, the categories of the framework and the text type, but it also pointed to methodological issues discussed in relation to our results and to the literature. We believe that the annotation protocol and our annotations will be a good basis for future studies since there are no duplicates in our set (all tweets have only one label), and this will be helpful in the replicability of the study. Finally, our annotated data can also be used in computational tasks such as stance detection and classification tasks.

REFERENCES

AlDayel, Abeer and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management* 58/4: 102597. https://doi.org/10.1016/j.ipm.2021.102597

Alonso Ameida, Francisco. 2015. Introduction to stance language. *Research in Corpus Linguistics* 3: 1–5.

Artstein, Ron and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34/4: 555–596.

Berman, Ruth, Hrafnhildur Ragnarsdóttir and Sven Strömqvist. 2002. Discourse stance: Written and spoken language. *Written Language & Literacy* 5/2: 255–289.

Biber, Douglas. 2006. Stance in spoken and written university registers. *Journal of English for Academic Purposes* 5/2: 97–116.

Boyd, Danah, Scott Golder and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on *Twitter*. *Proceedings of the 43rd Hawaii International*

Conference on System Sciences. Washington: IEEE Computer Society, 1–10. https://ieeexplore.ieee.org/document/5428313

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20/1: 37–46.

Du Bois, John W. 2007. The stance triangle. In Robert Englebretson ed. *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*. Amsterdam: John Benjamins, 139–182.

Ekberg, Lena and Carita Paradis. 2009. Evidentiality in language and cognition. *Functions of Language* 16/1: 5–7.

Facchinetti, Roberta, Frank Palmer and Manfred Krug. 2003. *Modality in Contemporary English*. Berlin: Walter de Gruyter.

Faulkner, Adam. 2014. Automated classification of stance in student essays: An approach using stance target information and the Wikipedia link-based measure. In William Eberle and Chutima Boonthum-Denecke eds. *Proceedings of the 27th International Florida Artificial Intelligence Research Society Conference*. Florida: Association for the Advancement of Artificial Intelligence, 174–179.

Feinstein, Alvan R. and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* 43/6: 543–549.

Ferreira, William and Andreas Vlachos. 2016. Emergent: A novel data-set for stance classification. In Kevin Knight, Ani Nenkova and Owen Rambow eds. *Proceedings of the Association for Computational Linguistics: Human Language Technologies*, 1163–1168. https://aclanthology.org/N16-1138/

Fuoli, Matteo. 2018. A stepwise method for annotating APPRAISAL. *Functions of Language* 25/2: 229–258.

Ghosh, Shalmoli, Prajwal Singhania, Siddharth Singh, Koustav Rudra and Saptarshi Ghosh. 2019. Stance detection in web and social media: A comparative study. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric SanJuan, Linda Cappellato and Nicola Ferro eds. *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*. Cham: Springer, 75–87.

Gwet, Kilem. 2002. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-rater Reliability Assessment* 1: 1–5.

Hasan, Kazi Saidul and Vincent Ng. 2014. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In Alessandro Moschitti, Bo Pang and Walter Daelemans eds. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha: Association for Computational Linguistics, 751–762.

Hernández, Nuria. 2014. New media, new challenges: Exploring the frontiers of corpus linguistics in the linguistics curriculum. *Research in Corpus Linguistics* 1: 17–31.

Hidalgo-Downing, Laura. 2012. Grammar and evaluation. *The Encyclopedia of Applied Linguistics*. https://doi.org/10.1002/9781405198431.wbeal1471

Hoek, Jet and Merel Scholman. 2017. Evaluating discourse annotation: Some recent insights and new approaches. In Harry Bunt ed. *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation*. Tilburg: Tilburg University, 1–13. https://aclanthology.org/W17-7401/

Honey, Courtenay and Susan C. Herring. 2009. Beyond microblogging: Conversation and collaboration via *Twitter*. *Proceedings of the 42nd Hawaii International Conference on System Sciences*. Waikoloa: IEEE Computer Society, 1–10. https://ieeexplore.ieee.org/document/4755499

Hripcsak, George and Daniel F. Heitjan. 2002. Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics* 35/2: 99–110.

Hunston, Susan and Geoffrey Thompson. 2000. *Evaluation in Text: Authorial Stance and the Construction of Discourse.* Oxford: Oxford University Press.

Hyland, Ken. 2005. Stance and engagement: A model of interaction in academic discourse. *Discourse Studies* 7/2: 173–192.

Jacknick, Christine M. and Sharon Avni. 2017. Shalom, bitches: Epistemic stance and identity work in an anonymous online forum. *Discourse, Context & Media* 15: 54–64.

Jaffe, Alexandra. 2009. *Stance: Sociolinguistic Perspectives*. Oxford: Oxford University Press.

Kaltenböck, Gunther, María José López-Couso and Belén Méndez-Naya. 2020. The dynamics of stance constructions. *Language Sciences* 82: 101330. https://doi.org/10.1016/j.langsci.2020.101330

Krippendorff, Klaus. 2011. *Computing Krippendorff's Alpha-reliability*. https://repository.upenn.edu/asc_papers/43. (24 November, 2022.)

Kucher, Kostiantyn, Andreas Kerren, Carita Paradis and Magnus Sahlgren. 2016. Visual analysis of text annotations for stance classification with ALVA. In Tobias Isenberg and Filip Sadlo eds. *Proccedings of the Eurographics Conference on Vizualization*, 49–51. http://dx.doi.org/10.2312/eurp.20161139

Küçük, Dilek and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys* 53/1: 1–37.

Landis, J. Richard and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33/1: 159–174.

Lombard, Matthew, Jennifer Snyder-Duch and Cheryl Campanella Bracken. 2002. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research* 28/4: 587–604.

Marín-Arrese, Juana I. 2017. Stancetaking and inter/subjectivity in journalistic discourse: The engagement system revisited. In Ruth Breeze and Inés Olza eds. *Evaluation in Media Discourse: European Perspectives*. Bern: Peter Lang, 21–48.

Marín-Arrese, Juana I., Marta Carretero, Jorge Arús Hita and Johan Van der Auwera eds. 2014. *English Modality: Core, Periphery and Evidentiality*. Berlin: Mouton de Gruyter.

McHugh, Mary L. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica* 22/3: 276–282.

Mohammad, Saif, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In Steven Bethard, Marine Carpuat, Daniel Cer, David Jurgens, Preslav Nakov and Tortsten Zesch eds. *Proceedings of the 10th International Workshop on Semantic Evaluation*. San Diego: Association for Computational Linguistics, 31–41. https://aclanthology.org/S16-1003/

Neuendorf, Kimberly. 2017. *The Content Analysis Guidebook*. Thousand Oaks: SAGE publications.

Pamungkas, Endang Wahyu, Valerio Basile and Viviana Patti. 2019. Stance classification for rumour analysis in *Twitter*: Exploiting affective information and conversation structure. *arXiv preprint arXiv: 1901.01911*. https://doi.org/10.48550/arXiv.1901.01911

Paradis, Carita. 2003. Between epistemic modality and degree: The case of *really*. *Topics in English Linguistics* 44: 191–222.

Perrin, Daniel. 2012. Stancing: Strategies of entextualizing stance in newswriting. *Discourse, Context & Media* 1/2–3: 135–147.

Poesio, Massimo. 2004. Discourse annotation and semantic annotation in the GNOME corpus. In Bonnie Webber and Donnna Byron eds. *Proceedings of the Workshop on Discourse Annotation*. Barcelona: Association for Computational Linguistics, 72–79.

Schmidt, Jan-Hinrik. 2014. Twitter and the rise of personal publics. In Katrin Weller, Alex Bruns, Jean Burgess, Merja Mahrt and Cornelius Puschmann eds. *Twitter and Society*. Bern: Peter Lang, 3–14.

Simaki, Vasiliki, Carita Paradis, Panagiotis Simakis and Andreas Kerren. 2017a. Stance classification in texts from blogs on the 2016 British referendum. In Alexey Karpov, Rodmonga Potapova and Losif Mporas eds. *Proceedings of the 19th Speech and Computer International Conference.* Charm: Springer, 700–709.

Simaki, Vasiliki, Carita Paradis and Andreas Kerren. 2017b. Identifying the authors' national variety of English in social media texts. In Ruslan Mitokov and Galia Angelova eds. *Proceedings of the Recent Advances in Natural Language Processing Conference*, 700–709. https://acl-bg.org/proceedings/2017/RANLP%202017/pdf/RANLP086.pdf

Simaki, Vasiliki, Carita Paradis and Andreas Kerren. 2018a. Evaluating stance-annotated sentences from the *Brexit Blog Corpus*: A quantitative linguistic analysis. *ICAME Journal* 42: 133–165.

Simaki, Vasiliki, Panagiotis Simakis, Carita Paradis and Andreas Kerren. 2018b. Detection of stance-related characteristics in social media text. In Nikos Fakotakis and Vasileios Megalooikonomou eds. *Proceeding of the 10th Hellenic Conference on Artificial Intelligence*. Patras: Association for Computing Machinery, 1–7. https://doi.org/10.1145/3200947.3201017

Simaki, Vasiliki, Carita Paradis and Andreas Kerren. 2019. A two-step procedure to identify lexical elements of stance constructions in discourse from political blogs. *Corpora* 14/3: 379–405.

Simaki, Vasiliki, Carita Paradis, Maria Skeppstedt, Magnus Sahlgren, Kostiantyn Kucher and Andreas Kerren. 2020. Annotating speaker stance in discourse: The *Brexit Blog Corpus*. *Corpus Linguistics and Linguistic Theory* 16/2: 215–248.

Taboada, Maite. 2016. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics* 2: 325–347.

Traugott, Elizabeth Closs. 2020. Expressions of stance-to-text: Discourse management markers as stance markers. *Language Sciences* 82: 101329. https://doi.org/10.1016/j.langsci.2020.101329

Verhagen, Arie. 2005. *Constructions of Intersubjectivity: Discourse, Syntax, and Cognition*. Oxford: Oxford University Press.

Yus, Francisco. 2011. *Cyberpragmatics: Internet-mediated Communication in Context*. Amsterdan: John Benjamins.

Yus, Francisco. 2016. Discourse, contextualization and identity shaping: The case of social networking sites and virtual worlds. In María Luisa Carrió-Pastor ed. *Technology Implementation in Second Language Teaching and Translation Studies*. Singapore: Springer, 71–88.

Zappavigna, Michele. 2012. *Discourse of Twitter and Social Media: How we Use Language to Create Affiliation on the Web*. London: A&C Black.

Zappavigna, Michele. 2015. Searchable talk: The linguistic functions of hashtags. *Social Semiotics* 25/3: 274–291.

Zappavigna, Michele and James R. Martin. 2018. # Communing affiliation: Social tagging as a resource for aligning around values in social media. *Discourse, Context & Media* 22: 4–12.
Zhu, Hongqiang. 2016. Searchable talk as discourse practice on the internet: The case of "# bindersfullofwomen." *Discourse, Context & Media* 12: 87–98.

Corresponding author
Vasiliki Simaki
Lund University
Faculties of Humanities and Theology
Centre for Languages and Literature
Helgonabacken 12
Box 201, SE 221 00
Lund
Sweden
E-mail: vasiliki.simaki@englund.lu.se

received: April 2022
accepted: November 2022
published online: December 2022

APPENDICES

Appendix 1: The framework's text stance categories in alphabetical order, followed by a brief description and examples (Simaki *et al.* 2020).

Stance category	**Description**	**Examples of utterances**
Agreement/ disagreement	The speaker expresses a similar or different opinion.	*I couldn't agree more to what you are saying.* *No, please don't do that.*
Certainty	The speaker expresses confidence as to what she or he is saying	*I am sure they will fight about it.* *Of course it is true.*
Contrariety	The speaker expresses a compromising or a contrastive/comparative opinion.	*While these are kind of notes to myself, you might still find them useful.* *The result is fairly good, but it could be better.*
Hypotheticality	The speaker expresses a possible consequence of a condition.	*If it's nice tomorrow, we will go.* *I will be happy, if Mike visits Granny tomorrow.*
Necessity	The speaker expresses a request, recommendation, instruction or an obligation.	*I must hand back all the books by tomorrow.* *This wine should drink well for two more decades.*
Prediction	The speaker expresses a guess/conjecture about a future event or an event in the future of the past.	*My guess is that the guests have already arrived.* *The meeting should not last longer than 2 hours.* *That ought to be fine.*
Source of knowledge	The speaker expresses the origin of what he or she says.	*I saw Mary talking to Elena yesterday.* *According to the news, the rate of interest is not going up.*
Tact/rudeness	The speaker expresses pleasantries and unpleasantries.	*Please, do give my love to him.* *You lazy bastard. Get lost.*
Uncertainty	The speaker expresses doubt as to the likelihood or truth of what she or he is saying.	*We have enough time, haven't we?* *There might be a few things left to do.*
Volition	The speaker expresses wishes or refusals, inclinations of disinclinations.	*I wish I could join you next summer.* *I prefer to stay in a cheap hotel.*

Appendix 2: Full list of stance markers for each stance category based on Simaki *et al.* (2019).

Contrariety	Hypotheticality	Necessity	Prediction	Source of knowledge	Uncertainty
And	*A*	*Have*	*Be*	*As*	*Could*
But	*Be*	*Let*	*May*	*Has*	*I*
Not	*Could*	*Must*	*Not*	*I*	*May*
Than	*If*	*Need*	*Is*	*Said*	*Maybe*
While	*In*	*Needs*	*It*	*Show*	*Might*
	Then	*Should*	*The*	*That*	*Probably*
	Will	*To*	*To*	*The*	*Think*
	Would	*We*	*Will*	*To*	

RiCL Research in Corpus Linguistics

Lexical indicators of profit and loss in Spanish shareholder letters

Blanca Carbajo Coronado
Autonomous University of Madrid / Spain

Abstract – Through a case, frequency, and collocational study, this work aims to detect linguistic differences between the letters to shareholders of profitable and loss-making companies. In a sample of 50 letters from each group, two lexical categories were analysed, verbs and eventive nouns, using the corpus manager *Sketch Engine*. The results indicated that verbs and eventive nouns with an absolute frequency of at least five occurrences overlap in both corpora by 95.7 per cent and 95.8 per cent, respectively. The frequency analysis showed that those with significantly higher frequency in one corpus denoted events or activities that were to be expected for companies in their group, such as *aumentar* ('increase'), *crecimiento* ('growth'), or *cumplimiento* ('compliance'), in the profitable companies; but terms such as *relanzar* ('relaunch'), *revalorización* ('revaluation'), or *pérdida* ('loss') in the loss-making companies. The analysis of the combinatorial properties of these verbs and nouns revealed subtle but significant differences between the two groups. In the case of verbs, the choice of the direct object is key, and in the case of nouns, qualifying and adverbial adjectives are crucial, as well as the Theme complements.

Keywords – finance; lexicon; collocations; verbs; eventive nouns

1. INTRODUCTION[1]

Companies publish information periodically on their initiatives or following regulations to enable potential investors and shareholders to assess their results, financial situation, and business plans. A shareholder letter is usually prepared once a year and included at the beginning of the company's annual report. The company's senior executives send this to their shareholders to provide a general summary of its operations for the entire year. The letter discusses the company's key financial results, market position, objectives, and

[1] This publication is part of the project *Computational linguistic methods for the readability and simplification of financial narratives. CLARA-FINT* (PID2020-116001RB-C31), funded by the Spanish Ministry of Science and Innovation and the State Research Agency. The author acknowledges the financial support provided by the FPU grant (FPU20/04007) which has been awarded by the Spanish Ministry of Science, Innovation and Universities.

business approaches. Additionally, specific events of the past year and share price fluctuations may also be stated.

The starting point of this paper is the assumption made by experts in the field of financial communication, who point out that loss-making companies mimic the themes and causality patterns of profitable companies (see Aerts 2005). Other works, such as the computational test by El-Haj *et al.* (2021), can lead to assuming a particular affinity or closeness between both groups. Based on machine learning, that study aimed at developing an automatic classifier of letters in profit or loss. Between 85 per cent and 89 per cent of the letters were correctly classified, but this is a low figure for a binary classification task in artificial intelligence.

This paper seeks to contribute to the study of shareholder letters with the help of the tools and methods of corpus linguistics in order to shed light on the usage and collocations of some lexical items, and to provide an insightful discussion on the language of shareholders' communications. Results can be valuable to linguists, especially working in the field of Spanish for specific purposes, as well as corporate communicators or business communication researchers.

A selection of 100 letters to shareholders are analysed, which are part of a corpus of Spanish annual reports called *FinT-esp* (Moreno-Sandoval *et al.* 2020). The analysis is based on the distinction between two groups of companies: those with profits and those with losses in the year under review. Little research has been carried out with the use of this corpus with the exception of García Toro (2020) on discourse markers. The distinction between companies is essential because, through the letters, shareholders may be able to detect any subtle changes in the company's strategy and learn where the company's senior executives stand on interpreting the results. Some companies may speak openly and transparently about their accounting results, but that is not true for all. In the letters, the results are either truthfully shown or hidden, depending on whether the company has made a profit or registered losses in the previous accounting year (Patelli and Pedrini 2014).

The focus of this paper is on indicators of profit or loss in companies. An 'indicator' is a loose term that refers to lexical items that are distinctive of either profit-making or loss-making companies, as they appear more frequently in one of both corpora. Ideally, lexical items will account for unique or almost unique themes to that group of companies. For example, a clear indicator would be *reparto (de dividendos)* ('(dividend) payout').

Only profitable companies own the required capital to pay out dividends to shareholders. By identifying such indicators, this study aims to provide insights into the linguistic features that distinguish profitable companies from their loss-making counterparts.

The categories chosen for the analysis are verbs and eventive nouns, that is, nouns designating events or something that happens within a period. The reason for this choice is that they account for the letters' most relevant themes. Action verbs, that is, activities, achievements, and accomplishments in Vendler's (1957) terminology (e.g. *competir* 'compete', *delegar* 'delegate', *implementar* 'implement') can shed light on the differences between the profit and loss groups in terms of the initiatives undertaken by the company during the respective financial year. Causative verbs such as *permitir* ('allow') or *provocar* ('provoke') can help understand the argumentation and justification of the company's decisions. They give an account of the point of view taken by companies to explain the events that occurred during the financial year. For example, the reduction in revenue may *have been caused* by a fiscal adjustment beyond the company's control. Another example is that the hard work of employees could be what *has enabled* record highs in revenue to be achieved. Verbs of change of state such as *reducir* ('reduce') and *crecer* ('grow') can give clues as to how the business is doing. On the other hand, copulative verbs, usually *ser* ('be') introduce information about how the year is evaluated in static terms by the company or the senior executive. Like verbs, nouns identify the critical events of the year and those that are expected to occur in the future. For example, one can talk about the *rise* in steel prices, in which the eventive noun identifies the main topic.

With the abovementioned information in mind, this paper aims to compare 50 letters from loss-making companies with 50 letters from profit-making companies to identify the differences between the texts of both groups regarding lexical aspects. More specifically, the research questions addressed are the following:

Q1) Are there any noteworthy lexical differences that set them apart?

Q2) In cases where the letters from both corpora employ the same lexical choices, what methods can be used to disambiguate them?

The paper is structured as follows. Section 2 provides an overview of the work done in financial communication studies and linguistics on letters to shareholders. It also introduces the characteristics of the letters, their communicative purposes, and the topics

they address. Section 3 describes the corpus of letters and the methodology. Section 4 presents the data analysis structured around the two lexical categories chosen. First, the differences between profit and loss letters in the frequency of use are discussed (Section 4.1. for verbs and 4.3. for eventive nouns). Then, the collocations that can help to disambiguate between the two groups of letters are studied (Section 4.2. for verbs and 4.4. for eventive nouns). The conclusions and final remarks are included in Section 5.

2. LETTERS TO SHAREHOLDERS

Letters to shareholders present the factual information contained in the annual report and assess the company's current situation, past performance, and growth potential. They constitute, therefore, a textual genre with an apparent persuasive character, not only because of the topics they cover but also because they explicitly identify their readers; their authors subscribe to what is said and sometimes state their position on the issues (Vogel 2020). These characteristics make letters to shareholders unique compared to other genres in the business world, which is one of the reasons why they have attracted a great deal of attention from experts in corporate investor relations.

Notable works have focused on the letters' move analysis, that is, the small semantic units with communicative purpose into which a text is divided (see Swales 1990; Garzone 2005; Bhatia 2008, among others). In Bhatia's (2008) proposal, a letter to shareholders contains the following movements: 1) reviewing the financial year, 2) identifying important issues, 3) developing those issues, 4) describing expectations and making promises, 5) expressing gratitude to employees and shareholders (optional), 6) talking about the future, and 7) closing positively and politely. Other authors propose variants, but these always include at least three movements: contextualisation/assessment of the year under discussion, description and justification of the financial year's performance, and discussion of future performance and strategy.

Some of those works focusing on moves address the topic from a comparative perspective. These include Garzone (2004), Nickerson and Groot (2005), Ruiz-Garrido *et al*. (2012), and Skorczynska and Giménez-Moreno (2016), among others. The work of Skorczynska and Giménez-Moreno (2016) reveals interesting characteristics of Spanish letters to shareholders, as opposed to British and Polish ones. For example, the fact that they share with the British ones the highest number of moves (a total of 15 moves

compared to 17 for the British and nine for the Polish ones), or that they reveal a preference for providing a great deal of information on the objectives, performance, strategy, investment, and work of the board of directors.

A number of studies have been particularly interested in the mechanisms of evaluation, which is understood as the expression of the speaker's or writer's attitude, or point of view towards the entities or propositions they are addressing (see Thompson and Hunston 2000), as well as the linguistic elements that make them explicit: epistemic modality markers (Kranich 2011; Kranich and Bicsar 2012), and adjectives or collocations (Poole 2016, 2017; Wang 2020; Skorczynska Sznajde 2021).

Adjectives and other evaluative categories, such as adverbs, can help determine how a company or senior executive perceives a particular event, action, or state. As a result, researchers in both linguistics and corporate communication studies have paid considerable attention to them. However, other lexical categories (primarily nouns or verbs) have received more attention in financial research, where there is a line of works focused on form-oriented content analysis, that is, the type and frequency of lexical units used (see Laskin 2018). This paper continues this line of research and provides new linguistic insights into these categories.

3. METHODOLOGY

A total of 100 letters of 24 companies from the period 2014–2018 were extracted from the *FinT-esp* corpus. All these letters belong to reports of companies listed on the Madrid Stock Exchange General Index and are distributed across eight different sectors following NACE,[2] the industry standard classification system of the European Union (see Table 1). The classification into profit and loss was carried out with the help of an expert in the field of finance, although the results of the year are available in the annual accounts contained in each company's report.

[2] https://nacev2.com/

Business activities	Profit corpus companies	Token count	%	Loss corpus companies	Token count	%
Mining and quarrying	Repsol	2,000	3.13	Repsol	1,786	2.85
Manufacturing	Gamesa, Acerinox, Ebro Foods, Almirall, Tubacex, Pharma Mar, Ercros, PRISA, Duro Felguera, Adveo, Adolfo Domínguez	40,808	63.79	Gamesa, Acerinox, Almirall, Deoleo, Tubacex, Pharma Mar, Ercros, PRISA, Duro Felguera, Adveo, Adolfo Domínguez	37,292	59.53
Electricity, gas, steam, and air conditioning supply	Gas Natural Fenosa	1,002	1.57	Naturgy Gas Natural	1,226	1.96
Water supply, sewerage, waste management, and remediation activities	Fluidra	2,392	3.74	Fluidra	2,606	4.16
Construction	Ferrovial, FCC, Sacyr	8,649	13.52	Ferrovial, FCC, Sacyr	12,823	20.49
Accommodation and food service activities	NH	1,756	2.75	NH	1,503	2.40
Information and communication	Telefónica, Indra	5,136	8.03	Indra	3,021	4.82
Real estate activities	Realia, Urbas	2,229	3.48	Realia, Urbas	2,390	3.82
Total		**63,972**			**62,647**	

Table 1: Companies in each business activity with their token count and percentage of words in the corpus

The underlying idea of the selection was to avoid biased results due to imbalances in the distribution of sectors in the corpora. The choice was conditioned from the outset by the number of loss letters available (60 vs. 332 for earnings), so the distribution of companies in the loss corpus was emulated in the profit corpus, with preference to the letters that belonged to the same company. Although the original *FinT-esp* corpus contained letters from 15 different business sectors, only eight were chosen for this study since only eight of them had companies incurring losses.

As the corpus includes letters from different companies, the results may be affected by differences between industries. Each sector displays a specific lexicon and tone in their corporate communications. For example, a mining company may use more technical language related to exploration, excavation, and extraction. In contrast, a real estate company may use more terms related to development, leasing, and sales. Additionally, the tone may also differ depending on the sector. For example, letters from companies in

the information and communication sector may be more upbeat and optimistic, while those in the mining and quarrying sector may be more cautious and focused on risk management. Luo and Zhou (2019) found that in industries related to oil and gas, tobacco, alcohol, or firearms less optimistic tones are used in order to avoid regulatory and public scrutiny. Despite focusing on detecting differences between company letters, this study recognises a marked similarity at the baseline between both corpora. Focusing on the differences between companies may overlook the similarities —in this case, the lexical ones— which are much more pronounced, as the literature reveals. Furthermore, the fact that the corpus tool chosen for this study is designed to look for differences may lead to bias when analysing data (Taylor 2013, 2018).

Sketch Engine (Kilgarrif *et al.* 2014) was the corpus manager tool chosen to conduct the lexical analysis. The texts were uploaded and subsequently divided into two sub-corpora: profit letters, which contained 63,972 tokens distributed across 1,823 sentences, and loss letters, with 62,647 tokens distributed across 2,069 sentences.

The lists of noun and verb lemmas from both sub-corpora were drawn using the *Word List* tool of *Sketch Engine*. The analysis was limited to representative lexical units in terms of frequency. Only the lemmas that appeared in the corresponding sub-corpus with an absolute frequency equal to or higher than five were selected. Admittedly, below this frequency, a large number of verbs and nouns can be found which show the greatest variation between the sub-corpora, but their low frequency prevent their consideration as prototypical units, which is the focus of the present work.

Subsequently, the lists from both sub-corpora were compared to obtain data for the same lemmas in both groups. Finally, in the case of eventive nouns, a manual selection was conducted based on the criteria for this lexical category described for Spanish in Bosque (1999). Eventive nouns share two key distributional properties: 1) they designate entities with temporal limits, which is why they are usually combined with verbs such as *empezar* (‘start’), *comenzar* (‘begin’), or *concluir* (‘conclude’), and 2) they take the preposition *durante* (‘during’) as their object.

Cases of ambiguity were found in those nouns where there is an eventive and an objectual interpretation, such as *gobierno* (‘government’), *consejo* (‘council’), or *negocio* (‘business’). Another type of ambiguity arose when there was an eventive and resultative interpretation in nouns, such as *beneficio* (‘profit’), *pérdida* (‘loss’), or *ganancia* (‘gain’). In the first interpretation, the referent is an event, as illustrated in (1). However, in the

second interpretation, the referent is the effect of the event designated by the verbal base of the noun, *perder* (‘lose’). Disambiguation occurs within the predicates where the noun phrase is located, as in (2). Due to their thematic relevance and the increased frequency of the eventive interpretation, these ambiguous cases were not discarded.

(1) Se produjo la **pérdida** del mercado asiático.
‘The loss of the Asian market occurred.’

(2) Tuvo unas **pérdidas** de 61,9 millones de euros.
‘It made a loss of 61.9 million euros.’

Following this process, a total of 345 verbs and 165 eventive nouns were retrieved, which are analysed in Section 4 with regard to their frequency and collocations. For the frequency comparison between the sub-corpora, the normalised frequency per thousand tokens has been adopted. Moreover, in cases where a particular unit occurs in both sub-corpora, the log-likelihood ratio has been calculated with Paul Rayson’s UCREL log-likelihood calculator[3] to determine whether the difference in frequencies is statistically significant. All those units that do not exceed 3.84 (p<0.05) were discarded (McEnery and Xiao 2005).

4. Results and analysis

The analysis of the data is divided into two parts. First, the differences between the two corpora regarding the frequency of verbs and eventive nouns are examined. The focus is primarily on those verbs and nouns that may be key to differentiate between the two groups of letters according to their frequency of use: firstly, those that do not appear in one of the two corpora and, secondly, those in both groups of letters, but in different proportions. On the other hand, the second part of the analysis deals with collocations. The purpose is to determine whether the more immediate context of verbs and nouns is relevant in those cases where the frequency of use does not allow significant differences to be drawn.

[3] https://ucrel.lancs.ac.uk/llwizard.html

4.1. Frequency analysis of verbs

A total 345 verbs were extracted, all used with a minimum absolute frequency of five in at least one of the two corpora. The data reveal that both corpora of letters use practically the same verbs. Only seven verbs which appear in the gain corpus are not present in the loss corpus: *colaborar* (‘collaborate’), *comportar* (‘involve’), *estabilizar* (‘stabilise’), *moderar* (‘reduce’), *optar* (‘opt for’), *propiciar* (‘favour’), and *sustituir* (‘replace’). In contrast, eight verbs from the loss corpus do not appear in the gain corpus: *centralizar* (‘centralise’), *coincidir* (‘coincide’), *combinar* (‘combine’), *detallar* (‘detail’), *indicar* (‘show’), *pertenecer* (‘belong’), *poseer* (‘own’), and *relanzar* (‘relaunch’). Thus, there is an overlap of 95.7 per cent between the sets of verbs in both corpora. This lexical similarity is one of the main reasons why it is challenging to classify the letters into two categories of companies, especially automatically.

In the following, there is a discussion on the use of verbs relevant for the differentiation between the two corpora. The term ‘relevant’ here implies lexical units used exclusively or almost exclusively by one of the two groups of companies, because of their semantic content and collocations. Therefore, if these lexical units appeared in a particular letter and we did not know what group they belonged to, they would be decisive in determining whether it is one group or the other. As can be seen, although the first selection is statistical —the focus is on units that do not occur in the other sub-corpus or with a difference in use statistically significant— there is a second selection of a subjective nature carried out after an in-depth examination of the letters. Therefore, all those verbs for which there is no further discussion are not to be considered relevant for the differentiation of both corpora.

4.1.1. Verbs exclusive to profit and loss-making companies

Three verbs that are only attested in one of the two groups of letters are analysed below. Their frequencies are shown in Table 2. In the following subsections, they are sorted alphabetically.

Verbs	Normalised frequency[4] in profit corpus	Normalised frequency in loss corpus
Optar 'opt for'	0.11	0
Propiciar 'favour'	0.13	0
Relanzar 'relaunch'	0	0.08

Table 2: Frequency of selected verbs exclusive to one corpus

4.1.1.1. *Optar* ('opt for')

Optar is frequently used by profit-making companies because of its prepositional regime complements introduced by *por* ('by'), such as *pago* ('payment'), *modalidad* ('modality'), or *efectivo* ('cash'), as in example (3), which are related to shareholders' decisions about the dividend they will receive. It is worth noting that it is not common for loss-making companies to pay out dividends.

(3) Nos permitió hacer inversiones por 156 millones de euros, atender el pago del dividendo de aquellos accionistas que **optaron** por efectivo en el "scrip dividend" del ejercicio (…). (Acerinox, Chairman, 2016)

'It allowed us to make investments of 156 million euros to pay the dividend to those shareholders who opted for cash in the scrip dividend for the year (…).'

4.1.1.2. *Propiciar* ('favour')

Propiciar is a causative verb that usually implies a positive connotation. Profit-making companies sometimes use it to disguise their direct involvement in the events described (because its subject usually denotes events and not entities) while, at the same time, making it clear that these events have occurred thanks to the company's good policies, as in example (4). At other times, *propiciar* describes situations in which the behaviour of some aspect of the market has led to another event considered positive. The former is encoded by the syntactic subject and other sentence elements that usually refer to the agent or cause, such as *evolución* ('evolution') or *materias primas* ('raw materials'). The latter is encoded by direct objects in active sentences and passive subjects, such as *meta* ('goal'), *desarrollo* ('development'), *aumento* ('increase'), *subida* ('rise'), or *mejora* ('improvement').

[4] Normalised frequency per thousand tokens.

(4) La consecución de nuestro Plan Estratégico 2018 con año y medio de antelación, las inversiones realizadas a lo largo del año y el crecimiento de nuestra actividad en los mercados maduros **han propiciado** un crecimiento del valor. (Fluidra, Chairman, 2017)

'The achievement of our 2018 Strategic Plan a year and a half in advance, the investments made throughout the year, and the growth of our activity in mature markets have led to a growth in value.'

4.1.1.3. *Relanzar* ('relaunch')

Loss-making firms use *relanzar* when a previously promoted action, initiative, or project could not be carried out or was not as successful as expected. Direct objects include *negocio* ('business'), as in example (5), *proyecto* ('project'), *crecimiento* ('growth'), or even *compañía* ('company').

(5) A partir de ahora trabajaremos para **relanzar** el negocio y lograr una compañía cercana a la producción y orientada al consumidor (…). (Deoleo, CEO, 2016)

'From now on, we will work to relaunch the business and achieve a company close to production and consumer-oriented [...].'

4.1.2. Verbs with a higher frequency in profit and loss-making companies

Three verbs with a statistically significant higher frequency in one of the two groups of letters are analysed below. Table 3 displays their frequencies. In the following subsections, they are arranged in alphabetical order.

Verbs	**Normalised frequency in profit corpus**	**Normalised frequency in loss corpus**	**Log-likelihood ratio**
Aumentar 'increase'	0.70	0.37	6.79
Lograr 'achieve'	0.23	0.73	17.17
Registrar 'record'	0.23	0.45	4.27

Table 3: Frequency and log-likelihood ratio of selected verbs appearing in both corpora

4.1.2.1. *Aumentar* ('increase')

Aumentar is an interesting verb in terms of its distributional properties since, by itself, it cannot disambiguate the two groups of letters. In the profit letters, it occurs typically in past tenses and with direct objects having to do with quantitative aspects of the business, such as *producción* ('production'), *exportación* ('export'), or *consumo* ('consumption'), as well as with direct objects related to qualitative aspects of the business such as

valoración (‘valuation’), as seen in example (6), *confianza* (‘trust’), *esfuerzo* (‘effort’), and *valor* (‘value’). The subjects of this verb in the profit letters are typically *beneficio* (‘profit’), *ebit*, or *ebitda* (two different financial profit indicators). As shown, *aumentar* is generally associated with indicators that have improved in the fiscal year.

(6) La propuesta de valor dirigida a cumplir y superar las expectativas de los consumidores está permitiendo que **aumente** la valoración sobre los hoteles. (NH, CEO, 2015)

‘The value proposition aimed at meeting and exceeding consumer expectations is allowing the valuation of hotels to increase.’

4.1.2.2. *Lograr* (‘achieve’)

The essential complements for this verb are the direct objects. Loss-making companies may aim at *equilibrio* (‘break-even’) or *éxito* (‘success’) but not at *ahorro* (‘savings’) or *contratación* (‘recruitment’), as they typically prioritise more pressing objectives, such as *saneamiento* (‘company restructuring’), as in example (7), *renovación* (‘renewal’), *recuperación* (‘turnaround’), *estabilidad* (‘stability’), *eficiencia* (‘efficiency’), *reducción de costes* (‘cost reduction’), *endeudamiento equilibrado* (‘balanced indebtedness’), or *mejora* (‘improvement’), among other essential aspects. These concerns would explain the frequency of *lograr* in the loss corpus, as these companies engage with their shareholders for more purposes than the profit-making companies because they need to convince them to make a future improvement in their performance.

(7) La contención del gasto debe seguir siendo un pilar de gestión ineludible en el Grupo FCC para **lograr** el pleno saneamiento de la compañía, a través de iniciativas alineadas con el ajuste y la reducción presupuestaria. (FCC, CEO, 2016)

‘Cost containment must remain an unavoidable management pillar in the FCC Group in order to achieve the full restructuring of the company through initiatives aligned with budget adjustment and reduction.’

The direct objects of *lograr* in the profit letters are most commonly *ahorro* (‘savings’), as in example (8), *contratación* (‘recruitment’), *equilibrio* (‘balance’), and *éxito* (‘success’), among others.

(8) Hemos conseguido el 49 % de nuestro objetivo de **lograr** un ahorro recurrente anual de 50 millones de euros […]. (Acerinox, CEO, 2017)

‘We have achieved 49 % of our target to achieve annual recurring savings of 50 million euros [...].’

4.1.2.3. *Registrar* (‘record’)

Registrar usually occurs in the loss corpus combined with *pérdida* (‘loss’), as in example (9), *caída* (‘fall’), or *reducción* (‘reduction’) as its direct objects. However, in the profit corpus, it also appears with nouns of positive connotation as direct objects, such as *crecimiento* (‘growth’), *evolución* (‘evolution’), or *beneficio* (‘profit’).

(9) El resultado atribuible en 2016 **registró** una pérdida neta de -162 millones de euros (…). (FCC, CEO, 2016)

‘The attributable result in 2016 recorded a net loss of -162 million euros [...].’

4.1.3. Further discussion

Even though there is an overwhelming similarity in lexical choice, some minor differences have been found as regards the verbs used in the two corpora of letters. The most relevant verbs in profit letters reflect typical activities of profitable companies. For example, there is talk of how *las inversiones propician un crecimiento del valor* (‘investments lead to value growth’). Additionally, shareholders can *optar por el pago* (‘opt for payment’), meaning they can receive dividends. The letters also mention improvements in performance indicators and business margins. For example, the following events are discussed: *aumentar la producción / la exportación / el consumo / la valoración / la confianza* (‘increase production / export / consumption / valuation / confidence)’, as well as *aumentar el beneficio / ebit / ebitda* (‘increase profit / ebit / ebitda)’), or *registrar un crecimiento / evolución / beneficio* (‘register growth / evolution / profit’).

On the other hand, in letters from loss-making companies, the most frequent verbs refer to the closing of the fiscal year without profits or refer to the objectives they promise to fulfil in the future. They use *relanzar el proyecto / el negocio / el crecimiento / la compañía* (‘relaunch the project / business / growth / company’) and talk about *lograr equilibrio / éxito / saneamiento / estabilidad / eficiencia / mejora* (‘achieve (balance / success / sanitation / stability / efficiency / improvement’), or *registrar pérdidas* (‘register losses’). Specific indicators such as *ebit* or *ebitda* are not referenced.

4.2. Verb disambiguation through collocations

As seen in Section 4.1., the gain and loss corpora share 95.7 per cent of the verbs with a minimum absolute frequency of five. Therefore, one must resort to syntactic collocations to find clear differences between both corpora. To this purpose, the 50 most frequent verbs were selected in the two corpora (see Appendix 1) and their distributional properties were studied. In this section, three types of collocations need to be distinguished:

1. Collocations which are exclusive to one corpus and not semantically equivalent to the collocations in the other corpus. In these cases, the verb or noun in question is combined in each corpus with different lexical units whose meanings are also markedly different. For example, in the case of direct objects, one may note the difference between *mantener la exigencia* ('keep (the) demand') in the profit letters and *mantener la rentabilidad* ('keep (the) profitability') in the loss letters.

2. Those which are exclusive to one corpus, but semantically equivalent or close to the collocations in the other corpus. Different lexical units are used in each corpus, but their meaning is similar; that is, verbs and eventive nouns appear in similar contexts. For example, the collocation of *mejorar* ('improve') with *considerablemente* ('considerably') in the profit letters, and with *sustancialmente* ('substantially') in the loss letters, both adverbial modifiers of gradation or intensity with similar meanings.

3. Those which are identical in both corpora. For example, the adverb *intensamente* ('intensely') modifies *trabajar* ('work') in both corpora.

For our purpose, type (1) collocations are the only ones suitable to discern between the two groups of companies. The following is a description of their combinations, grouped according to their syntactic functions. The verbs are listed in alphabetical order.

4.2.1. *Alcanzar* ('reach')

Alcanzar combines with direct objects referring to scalar attributes and their values, which are vital to differentiate between the two groups of letters. In the case of the profitable companies, *récord* ('record'), *competitividad* ('competitiveness'), *beneficio* ('profit'), or *tasa* ('rate') are used, alluding to the significant advances (high or maximum values of

the respective dimensions) that have been achieved. By contrast, these values are usually lower in loss letters: *alcanzar el estándar* (‘reaching the standard’) or *alcanzar la expectativa* (‘reaching the expectation’).

4.2.2. *Conseguir* (‘achieve’)

This verb is usually attested in different syntactic structures in both types of letters. In the profit letters, *conseguir* is used with direct nominal objects referring to different profit indicators, such as *cota de excelencia* (‘excellence benchmark’), *ebitda*, *eficiencia* (‘efficiency’), or *crecimiento* (‘growth’). Although loss-making companies can point to their *ebitda*, this term will never be the direct object of *conseguir* in loss letters, as it is a verb with positive connotations and this indicator will be a negative figure. In the letters of loss-making companies, *conseguir* is usually followed by infinitive objects (in the terminology of *Sketch Engine*), such as *batir la tendencia general del mercado* (‘to beat the market general trend’) and *volver a la rentabilidad* (‘to return to profitability’).

4.2.3. *Marcar* (‘mark’)

Marcar is used in the profit corpus with complements referring to the positive events that have taken place during the financial year, although it is used with two different meanings. Taking the *Diccionario de la Lengua Española* (DLE 2022) as a reference, *marcar* is used in the sense ‘prescribe, determine, or fix’ with direct objects such as *cumplimiento* (‘compliance’) or with subjects in passive sentences such as *previsión* (‘forecast’). On the other hand, the ninth sense found in the DLE (‘divide spaces, with milestones or signs of any kind, or divide them mentally’) reflects the use of this verb when combined with nouns such as *hito* (‘milestone’), where this lexical unit acts as a subject in a passive structure.

Likewise, loss-making companies use *marcar* with this last meaning. It combines with direct objects such as *camino* (‘path’), *línea* (‘line’), or *diferencia* (‘difference’). Having failed to meet targets, companies argue that the year has been a turning point.

4.2.4. *Seguir* (‘continue’)

This verb is the head of the durative periphrasis *seguir* (‘keep’) + gerund. In the profit letters, the main verb refers to processes (often incremental) with a positive connotation, as in *cosechando* (‘earning’), *amortizando* (‘amortising’), *evolucionando* (‘evolving’), *enriqueciendo* (‘enriching’), *prosperando* (‘prospering’), *progresando* (‘progressing’), *perfeccionando* (‘perfecting’), *compitiendo* (‘competing’), *consolidando* (‘consolidating’), *afianzando* (‘strengthening’), or *incrementando* (‘increasing’). In the letters from loss-making companies, the main verb of the periphrasis alludes to difficulties that are being overcome or must be overcome, as in *fortaleciéndonos* (‘strengthening (ourselves)’) and *viviendo* (‘living’), to indicate that challenges have marked the year, or *ajustando* (‘adjusting’), which refers to costs.

4.2.5. *Superar* (‘exceed’)

In the sense ‘exceed a limit’, *superar* resembles *alcanzar* (‘reach’) in that it often combines with direct objects referring to scalar attributes and their values. In the profit letters, it appears with direct objects such as *récord* (‘record’), *índice* (‘index’), *cifra* (‘figure’), or *previsión* (‘forecast’), which denote specific limits. In the loss letters, it often serves a different meaning, that is, to ‘overcome obstacles or difficulties.’ In these letters, it is combined with objects such as *año* (‘year’), *crisis* (‘crisis’), and *dificultad* (‘difficulty’), referring to the intricacies of the year.

4.2.6. Further discussion

Some conclusions can be made drawn from what has been discussed in the previous subsections as regards the collocations of verbs in both corpora. The most relevant syntactic function for distinguishing between letters from profitable companies and those with losses is the direct object, which is used differently in both groups of letters. On the one hand, profitable companies use direct objects related to qualitative milestones (*cota de excelencia* ‘excellence threshold’, *hito* ‘milestone’, *liderazgo* ‘leadership’) and quantitative business issues (*records* ‘records’, *índices* ‘indices’), suggesting a concern for excellence and growth in different business indicators.

On the other hand, loss-making companies tend to mention qualitative milestones, as the quantitative ones are unfavourable — they have not achieved profitability. For example, they use *alcanzar el estándar / las expectativas* 'reach the standard / expectations', *conseguir la vuelta a la rentabilidad* 'achieve the return to profitability', *marcar la diferencia* 'make a difference' (in their sector), *seguir fortaleciéndonos* 'continue to strengthen (ourselves)', *superar una dificultad* 'overcome a difficulty'. They focus on the possibility of change and improvement, while profitable companies talk about the opportunity for growth and expansion.

4.3. Frequency analysis of eventive nouns

A total of 165 eventive nouns were detected in the list of noun lemmas with a minimum absolute frequency of five. The data reveal that practically the same nouns are used in both corpora. There are only three nouns exclusive to the profit letters (*globalización* 'globalisation', *protección* 'protection', and *laminación* 'lamination') and four nouns exclusive to the loss letters (*fraude* 'fraud', *negociación* 'negotiation', *recogida*, 'collection', and *revalorización* 'revaluation'), representing an overlap of 95.8 per cent between the nouns of both corpora. In the following, the distinctive nouns of each corpus are included, as well as their relevance in differentiating the two types of companies.

As explained in Section 4.1, the criterion for considering a unit as distinctive is primarily statistical: they are only attested in one the corpora or are more frequent in one of them, and this difference in frequency is statistically significant. In addition, there is an eminently subjective criterion; after reading and studying the letters, only those which, due to their semantic content, tip the balance in favour of one group or the other are selected. For example, *recogida* ('collection') only appears in the loss letters, but it refers to *waste collection*, so that it cannot be a relevant unit in the loss corpus since, in the profit group, there are also manufacturing companies that may refer to waste recycling. This consideration is not absolute; instead, there are units that may attested in both corpora but combine with other items that can be used to disambiguate between letters.

4.3.1. Nouns exclusive to profit and loss-making companies

Nouns that are only found in one of the two groups of letters are discussed below. Their frequencies are shown in Table 4. In the following subsections, they are sorted in alphabetical order.

Nouns	**Normalised frequency in profit corpus**	**Normalised frequency in loss corpus**
Globalización 'globalisation'	0.08	0
Negociación 'negotiation'	0	0.13
Protección 'protection'	0.09	0
Revalorización 'revaluation'	0	0.08

Table 4: Frequency of selected nouns exclusive to one corpus

4.3.1.1. *Globalización* ('globalisation')

This noun is used with a negative connotation in the profit letters. Companies are concerned about the globalisation of markets, which is perceived as a threat that companies must face, as illustrated in example (10).

(10) Ninguna otra empresa productora de acero inoxidable en el mundo goza de tan buena posición para afrontar el difícil proceso de **globalización** de la economía, en su nueva versión de "**globalización** con barreras" (…). (Acerinox, CEO 2016)
'No other stainless steel producing company in the world enjoys such a good position to face the difficult process of globalisation of the economy, in its new version of "globalisation with barriers" [...].'

4.3.1.2. *Negociación* ('negotiation')

The chairmen or chairwomen and CEOs of loss-making companies report the negotiations with banks to define debt interest rates. Therefore, *negociación* can be combined with *banca* ('banking') as the object of the preposition *con* ('with'), or with *refinanciación* ('refinancing'), the direct object (Theme) of the base verb *negociar* ('negotiate'), as in example (11).

(11) Este documento (…) es la base sobre la que se está apoyando la **negociación** con la banca cuyo primer objetivo es cerrar un "Acuerdo de Espera" o ("Standstill") para alcanzar posteriormente una reestructuración de la deuda. (Duro Felguera, Chairman, 2016)

> ‘This document [...] constitutes the basis for the negotiations with the banks. Its first objective is to conclude a ‘standstill’ in order to reach a debt restructuring subsequently.’

4.3.1.3. *Protección* (‘protection’)

Profit-making companies refer to the protection of three areas: investment (*protección de los intereses de los accionistas* ‘protection of shareholders’ interests’), environment (*protección del entorno ambiental* ‘protection of the environment’), and labour (*protección y seguridad de su personal expatriado* ‘protection and safety of their expatriate staff’).

4.3.1.4. *Revalorización* (‘revaluation’)

Revalorización implies that an asset or security is at low levels at the time of speech. Letters from loss-making companies consider it relevant to comment on this (possible) increase in the value of their company’s stock, as illustrated in example (12). This rise is good news, implying that investors are confident in the business performance. Profit-making companies with good results do not need to appeal to investors’ confidence in this manner.

(12) Estamos convencidos de que los mercados no están reflejando adecuadamente el valor intrínseco de nuestro valor y que existe un importante potencial de **revalorización** de nuestra acción. (Acerinox, Chairman, 2018)

‘We are convinced that the markets are not adequately reflecting the intrinsic value of our stock and that there is significant upside potential for our share.’

4.3.2. Nouns with a higher frequency in profit and loss-making companies

The following is an analysis of eventive nouns that occur with a statistically significant higher frequency in one of the two corpora of letters. Table 5 displays their frequencies. In the subsections below, they are sorted alphabetically.

Nouns	Normalised frequencies in profit corpus	Normalised frequencies in loss corpus	Log-likelihood ratio
Ahorro 'saving'	0.17	0.37	4.58
Ampliación 'expansion'	0.22	0.61	12.02
Cumplimiento 'compliance'	0.42	0.19	5.61
Organización 'organisation'	0.17	0.35	3.97
Pérdida 'loss'	0.22	0.49	6.95
Reestructuración 'restructuring'	0.22	0.49	6.95

Table 5: Frequency and log-likelihood ratio of selected nouns appearing in both corpora

4.3.2.1. *Ahorro* ('saving')

Saving costs and funds are a critical factor for loss-making companies, which encourage *medidas de ahorro* ('savings measures'), *acciones de ahorro* ('savings actions'), or *planes de ahorro* ('savings plans'). For these companies, it is an obligation while for profitable companies, it is a supplement or an additional merit of their management. Therefore, instead of acting as a complement of nouns referring to different types of projects, *ahorro* usually appears as a noun head modified by the adjective in profit letters, as, for example, *ahorro adicional / recurrente* ('additional / recurrent savings'). It also functions as a direct object of verbs, such as *destacar* ('highlight') and *lograr* ('achieve'), with positive connotations.

4.3.2.2. *Ampliación* ('expansion')

Ampliaciones de capital ('capital increases') are frequently referred to in the loss letters. They are operations aimed at increasing the company's resources, which can be done by increasing the number of shares or their nominal value, that is, the value assigned by the owner. Although a profit-making company may also carry out capital increases to make investments, capital increases are typical of loss-making companies, as they allow them to meet their debts. Its occurrence should therefore tip the balance in favour of loss-making. When *ampliación* is modified by adjectives such as *segunda* ('second') or *última* ('last'), it becomes more evident that it belongs to the loss corpus. In the case of gains, texts mention the *ampliación de la capacidad* ('expansion of capacity'), referring to the increase in plant output in the case of manufacturing companies.

4.3.2.3. *Cumplimiento* (‘compliance’)

Profit-making companies refer to all the commitments and obligations they have managed to meet, be their environmental or governance policies, measures, laws, or business plans, as illustrated in example (13). There is even discussion of a ‘culture of compliance’ (*cultura del cumplimiento*), of which, for obvious reasons, loss-making companies cannot be a part of.

(13) Dará como resultado un Grupo más fuerte, más innovador, más competitivo y más comprometido con los valores tradicionales de nuestra compañía: la prudencia, la austeridad, la calidad y el **cumplimiento** de los compromisos asumidos. (Sacyr, Chairman, 2016)

‘It will result in a stronger, more innovative, more competitive Group that is more committed to the traditional values of our company: prudence, austerity, quality and compliance with our commitments.’

4.3.2.4. *Organización* (‘organisation’)

Loss letters refer to the different organisations they support and provide them with the necessary social prestige. Note, for example, (14), where there is an ambiguity between the objectual and eventive interpretations. Loss letters can also allude to the need for efficiency, as illustrated in (15). By contrast, profit letters praise their *cultura de organización* (‘organisational culture’).

(14) Fluidra ha reforzado significativamente su **organización** mundial de I+D. (Fluidra, Chairman, 2018)

‘Fluidra has significantly strengthened its worldwide R&D organisation.’

(15) En paralelo a estas medidas se ha definido una **organización** más eficiente (Deoleo, Chairman, 2016)

‘In parallel to these measures a more efficient organisation has been defined’.

4.3.2.5. *Pérdida* (‘loss’)

Unsurprisingly, *pérdidas* (‘losses’) are mentioned much more frequently in the loss letters than in the profit letters. These include *pérdidas contables* (‘accounting losses’), *pérdidas operativas* (‘operating losses’), *pérdidas de margen* (‘margin losses’), and *pérdidas de volumen* (‘volume losses’), among others. The profit letters, on the other hand, point out

losses *en las ventas* (‘on sales’), and highlight gains compared to losses in previous years or the value of the shares in falling markets, as illustrated in example (16).

(16) Se revalorizó en un 33,9% durante el ejercicio frente a la **pérdida** del 2,0% que sufrió el IBEX-35. (Acerinox, CEO, 2016)

‘It was revalued by 33.9% during the year compared to the 2.0% loss suffered by the IBEX-35.’

4.3.2.6. *Reestructuración* (‘restructuring’)

In the loss letters, *reestructuración* is a complement of nouns referring to a process, such as *actuación* (‘performance’) or a part of it, *culminación* (‘completion’), and to other nouns, such as *esfuerzo* (‘effort’), implying that it is a change forced by the situation. In loss-making companies, in turn, it also combines with *industrial* (‘industrial’), *logística* (‘logistical’), a classificatory relational adjective, *organizativa* (‘organisational’, where *organizativa* refers to the subject of the base verb) or *de plantilla* (‘workforce’), among others. Profit-making companies use these as well. However, unlike the loss-making companies, they refer to a ‘shareholding restructuring’ (*reestructuración accionarial*) or the restructuring of a ‘business division’ (*división del negocio*), in specific company areas.

4.3.3. Further discussion

The previous analysis revealed that the distinctive eventive nouns of each corpus depict activities typical of their group. Specific nouns, including *negociación* (‘negotiation’) and *revalorización* (‘revaluation’), exhibit significant differences between the two groups. Loss-making companies use them to refer to their debt management and the possibility of increasing the value of their shares in the market. These companies also focus on the actions they must take to achieve profits in the future (*planes de ahorro* ‘savings plans’, *ampliación de capital* ‘capital expansion’, *reestructuración logística / organizativa* ‘(logistics / organisational restructuring’). On the other hand, profitable companies focus on achieving previously established growth and objectives (*cumplimiento* ‘compliance’). Additionally, corporate social responsibility is an essential topic for both, as it can affect their social prestige and, therefore, their image before shareholders and the general public. In the case of loss-making companies, their letters reference the *organizaciones* (‘organisations’) they support.

4.4. Noun disambiguation through collocations

As discussed in Section 4.3, the profit and loss corpora share 95.8 per cent of the eventive nouns with a minimum absolute frequency of five in at least one corpus. This overlap means that, in most cases, to distinguish between the gain and the loss letters, one must resort to collocations. To this end, as in the case of verbs, the 50 most frequent eventive nouns in each corpus were selected (see Appendix 2).

As in the case of verbs (see Section 4.2), the focus will be placed on the collocations of type (1), the most suitable to differentiate between both groups of letters. It is worth recalling that type (1) collocations are exclusive to one corpus and bear no semantic resemblance to those of the other corpus. The nouns are presented below in alphabetical order.

4.4.1. *Acuerdo* (‘agreement’)

The profit-making companies use adverbial adjectives such as *definitivo* (‘definitive’) and evaluative adjectives such as *oportuno* (‘timely’) to define the agreements they sign. They also close *acuerdos de fusión* (‘merger deals’), which can be perceived as a business growth operation. Loss-making companies use evaluative qualifying adjectives with a negative connotation to define the agreements they sign, such as *malo* (‘bad’).

4.4.2. *Aumento* (‘increase)

Aumento (‘increase’) is a noun that introduces gradual events. Profit letters speak of *aumento significativo del volumen de inversión* (‘significant increase in the company’s investment volume’) or *aumento de la productividad de la compañía* (‘increase in the company’s productivity’) and, in other cases, allude to *aumento salarial* (‘wage increase’). Another indication of earnings is *aumento de dividendos* (‘increase in dividends’). It is an action that can be considered unequivocal of earnings because their distribution is not mandatory and their rise less so. To a lesser extent, *aumento de las ventas o de la demanda* (‘increase in sales or demand’) can also be distinctive of profit. In all these cases, the preposition object is the Theme argument of *aumento*.

4.4.3. *Comportamiento* (‘behaviour’)

The adjective modifiers in the profit letters make it clear that the behaviour has been good in its different aspects. *Comportamineto ético* (‘ethical behaviour’) is defined as *impecable* (‘impeccable’), *estrategia de comunicación* (‘communication strategy’) as *intachable* (‘faultless’), *nuevas gamas* (‘new ranges’) as *extraordinarias* (‘outstanding’), and the markets in which the company participates as *excelentes* (‘excellent’). The implementation of a *guía de comportamiento íntegro* (‘code of integrity’) is also mentioned. By contrast, loss letters prefer to use the comparative adjective *peor* (‘worse’) to assess the company’s performance, as in (17).

(17) Acerinox, que había evolucionado en línea con el Ibex 35 durante el primer semestre, tuvo un peor **comportamiento** que este índice en el cuarto trimestre (Acerinox, Chairman, 2018).

‘Acerinox, which had evolved in line with the Ibex 35 during the year’s first half, performed worse than this index in the fourth quarter.’

4.4.4. *Compromiso* (‘commitment’)

To define their commitment, profitable companies use evaluative adjectives that denote the highest degree of a property, such as *ejemplar* (‘model’), *altísimo* (‘very high’), and *mayor* (‘higher’), the latter used to compare the current year with past years. They also employ adjectives such as *constante* (‘constant’). On the other hand, the loss-making companies avoid qualifying adjectives and use relational adjectives such as *público* (‘public’) or adverbial adjectives such as *mutuo* (‘mutual’), referring to the commitment between the companies and the employees of the group.

4.4.5. *Crecimiento* (‘growth’)

In the profit letters, the modifiers of *crecimiento* are adjectives of degree and intensity, used as evaluative and adverbial modifiers, such as *fuerte* (‘strong’), *relevante* (‘relevant’), *mayor* (‘major’), *potente* (‘powerful’), *importante* (‘important’), *significativo* (‘significant’), *mejor* (‘better’), *notable* (‘remarkable’), *continuo* (‘continuous’), *sostenible* (‘sustainable’), *rentable* (‘profitable’), *sólido* (‘solid’), *firme* (‘steady’), or *exponencial* (‘exponential’), among others.

Loss letters use specific adjective modifiers. They judge their growth as *desigual* (‘uneven’) and promise to meet more *ambiciosos* (‘ambitious’) growth rates. Also, *crecimiento* (‘growth’) appears as a direct object of verbs such as *relanzar* (‘relaunch’), implying that growth has not occurred in past years.

4.4.6. *Ejercicio* (‘year’)

The adjectival modifiers of *ejercicio* are different in each corpus. In the profit letters, it is combined with evaluative qualifying adjectives such as *bueno* (‘good’). By contrast, in the loss letters, it occurs together with the adjectives *arriesgado* (‘risky’), *decisivo* (‘decisive’), or *complejo* (‘complex’), among others.

4.4.7. *Evolución* (‘evolution’)

The profitable companies describe *evolución de las ventas* (‘sales evolution’) with qualifying adjectives such as *alcista* (‘bullish’), with a descriptive interpretation, and *espectacular* (‘spectacular’), with an evaluative interpretation. On the other hand, the loss letters use evaluative modifiers to define the *evolución* of different aspects of their business as *desfavorable* (‘unfavourable’) and *peor* (‘worse’).

4.4.8. *Inversión* (‘investment’)

The profit letters refer to ‘investments’ *arranque* (‘beginning’) or *fomento* (‘promotion’), with investment playing the semantic role of Theme (effected in the first case and affected in the second). The quantitative dimension of investments is introduced by the phrases *grado de inversion* (‘investment degree’), *importe de inversion* (‘investment amount’), and *volumen de inversión* (‘investment volume’). Companies with losses mention their *reducción* (‘reduction’), *contracción* (‘contraction’), or *parálisis* (‘stoppage’) as part of nominal groups in which *inversión* is assigned the role of affected Theme.

4.4.9. *Mejora* (‘improvement’)

The profitable firms use evaluative adjectives such as *grandes* (‘great’) to refer to the technological improvements (*mejoras tecnológicas*) they have made and use the adverbial

adjective *constante* (‘constant’) to define them. By contrast, the loss-making companies prefer to use adjectives such as *paulatinas* (‘gradual’) or *inminentes* (‘imminent’).

4.4.10. *Proceso* (‘process’)

The profitable companies mention *procesos de producción* (‘production processes’). In the loss letters, *proceso* is combined with prepositional complements with nouns such as *refinanciación* (‘refinancing’) and *desinversión* (‘disinvestment’), which are symptoms of losses in a business.

4.4.11. *Reducción* (‘reduction’)

Like *aumento,* the noun *reducción* introduces gradual events. In the profit letters, it is accompanied by argument Themes such as *accidentes* (‘accidents’) or *accidentabilidad* (‘accidentability’), referring to production plants, and *emisiones de carbono* (‘carbon emissions) or *emisiones de gases de efecto invernadero* (‘greenhouse gas emissions’). In the letters of loss-making companies, its arguments include *producción* (‘production’), *capital* (‘capital’), or *inversiones* (‘investments’). In addition, loss-making companies refer to *reducción presupuestaria* (‘budget reduction’). Also, *reducción* is part of the prepositional complements of nouns such as *esfuerzo* (‘effort’), *programa* (‘programme’), *política* (‘policy’), or *cultura* (‘culture’), which are expressions that imply that the company is committed to reducing costs and debt.

4.4.12. *Result* (‘resultado’)

Adjective modifiers are vital for the disambiguation of *resultado,* especially those that are evaluative and qualifying and denote a high or maximum degree of a property. Profit letters assess the last fiscal year’s results as *positivos* (‘positive’), *buenos* (‘good’), *excelentes* (‘excellent’), *sólidos* (‘solid’), *magníficos* (‘great’), *extraordinarios* (‘extraordinary’), *significativos* (‘significant’), *grandes* (‘big’), or as *los mejores* ‘the best’. By contrast, in the loss letters, they are *inferiores* (‘lower’).

4.4.13. Further discussion

The analysis of collocations showed that adjectival modifiers, prepositional complements introduced by *de* and head of noun groups are the most important syntactic functions to distinguish between both corpora of letters. Moreover, profitable companies use evaluative adjectives with positive connotations to describe their performance (*crecimiento significativo* 'significant growth', *mejoras grandes* 'great improvements', *resultados excelentes* 'excellent results'). In contrast, loss-making companies use evaluative adjectives with negative connotations (*malos acuerdos* 'bad deals', *peor comportamiento* 'worse performance', *evolución desfavorable* 'unfavourable development', *resultados inferiores* 'inferior results'). Concerning other types of collocations, profitable companies discuss increases in various positive areas of business performance (*aumento del volumen de la inversión / de la productividad / de dividendos / salarial*, 'increase in investment volume / productivity / dividends / wages'), while loss-making companies talk about downsising and *procesos de desinversión / refinanciación* ('disinvestment/ financing processes').

5. Conclusion

This paper has analysed a selection of one hundred letters to shareholders written by companies listed on the Madrid Stock Exchange General Index between 2014–2018. The aim was to find lexical clues that would allow to differentiate between letters from profitable and letters from loss-making companies.

However, it is worth noticing that the present study has some limitations and contextualise the results accurately. First, the corpus size may not represent all letters to shareholders listed on the Madrid Stock Exchange General Index. Secondly, the study may have mainly focused on the differences between the corpora, neglecting the similarities that were clear from the outset. Additionally, the data selection and discussion methodology, which included a mixture of quantitative and subjective analysis, could have played a role on the results. Finally, focusing on two specific grammatical categories may not fully capture the complexity of the language used in these letters.

The initial hypotheses pointed to a lexical similarity in both corpora because the letters from the loss-making companies tended to mimic the profit-making companies' themes and causality patterns. These hypotheses have been confirmed, as the lexical

choice in both corpora is similar. In the grammatical categories analysed in this paper, that is, verbs and eventive nouns, there is an overlap of 95.7 per cent and 95.8 per cent, respectively.

The characteristic verbs that profitable companies make use of are those that report the increase in business margins or the distribution of dividends (*optar* 'opt for', *propiciar* 'propitiate', *aumentar* 'increase'). On the other hand, those related to closing the fiscal year without profits, meeting their objectives, or promising to fulfil them in the future (*relanzar* 'relaunch', *registrar* 'register') are characteristic of the loss-making companies.

As far as the nouns are concerned, those with a higher frequency in profitable companies refer to the achievement of the objectives proposed in previous years (*cumplimiento* 'compliance'). Similarly, the nouns with significantly more frequency in the loss-making companies refer to the losses themselves or the actions the company must take concerning products and business management to overcome those poor results: *ahorro* ('savings'), *ampliación* ('expansion'), *reestructuración* ('restructuring), *organización* ('organisation'), etc.

Furthermore, in the analysis of the distributional properties of verbs and eventive nouns, the study revealed that the syntactic element that showed the most prominent differences between the two groups is the direct object, rather than the subject or other arguments. The most noticeable difference in the collocations is that profitable companies can talk about qualitative achievements (the company is a leader in the sector, has proved to be a company that delivers, has the confidence of shareholders, and has managed well) and quantitative achievements (business margins are positive). By contrast, the loss-making companies can only mention qualitative achievements (e.g., the company's leading position) and negative quantitative data.

In the case of eventive nouns, the critical syntactic functions in the distinction between the letters are the adjectival modifiers, especially qualifying and adverbial, prepositional complements introduced by *de* (usually argumental with the role of Theme), and head in those cases where the noun in question acts as a complement to other nouns.

In short, despite the similarity between the two corpora, there are lexical signs that allow us to differentiate between them. However, such differences are very subtle because the linguistic reality of the letters is very complex. In fact, an examination carried out solely at the lexical or even morphosyntactic level will likely not reveal any real

significant differences. Therefore, it would be more productive to analyse the structure and argumentation patterns of the letters in depth, as well as the data that the companies choose to omit. These are central aspects for our future work, which will help us to understand the genre of letters to shareholders in Spanish better.

REFERENCES

Aerts, Walter. 2005. Picking up the pieces: Impression management in the retrospective attributional framing of accounting outcomes. *Accounting, Organizations and Society* 30: 493–517.

Bhatia, Vijay K. 2008. Towards critical genre analysis. In Vijay K. Bhatia, John Flowerdew and Rodney H. Jones eds. *Advances in Discourse Studies*. London: Routledge, 166–177.

Bosque, Ignacio. 1999. El nombre común. In Violeta Demonte and Ignacio Bosque eds. *Gramática Descriptiva de la Lengua Española*. Madrid: Espasa Calpe, 3–75.

El-Haj, Mahmoud, Antonio Moreno-Sandoval and José Antonio Jiménez Millán. 2021. Machine learning models for classifying Spanish beaters and non-beaters financial reports. In Antonio Moreno-Sandoval ed. *Financial Narrative Processing in Spanish*. Valencia: Tirant, 179–198.

García Toro, Ana. 2022. Los marcadores del discurso en la narrativa financiera: Análisis de las cartas a los accionistas. *ELUA. Estudios de Lingüística* 38: 187–214.

Garzone, Giuliana. 2004. Annual company reports and CEO's letters: Discoursal features and cultural markedness. In Christopher Candlin and Maurizio Gotti eds. *Intercultural Aspects of Specialized Discourse*. Bern: Peter Lang, 311–341.

Garzone, Giuliana. 2005. Letters to shareholders and chairman's statements: Textual variability and generic integrity. In Paul Gillaerts and Maurizio Gotti eds. *Genre Variation in Business Letters*. Bern: Peter Lang, 179–204.

Kranich, Svenja. 2011. The use of epistemic expressions in letters to shareholders in France, the United States and Germany. *Langage et Société* 3/3: 115–134.

Kranich, Svenja and Andrea Bicsar. 2012. "These forecasts may be substantially different from actual results". The use of epistemic modal markers in English and German original letters to shareholders and in English-German translations. *Linguistik Online* 55/5: 41–55.

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

Laskin, Alexander V. 2018. The narrative strategies of winners and losers: Analyzing annual reports of publicly traded corporations. *International Journal of Business Communication* 55/3: 338–356.

Luo, Yan and Linying Zhou. 2019. Tone of earnings announcements in sin industries. *Asian Review of Accounting* 27/2: 228–246.

McEnery, Anthony and Zhonghua Xiao. 2005. *Help* or *Help to*: What do corpora have to say? *English Studies* 86/2: 161-187.

Moreno-Sandoval, Antonio, Ana Gisbert and Helena Montoro. 2020. FinT-esp: A corpus of financial reports in Spanish. In Miguel Fuster-Márquez, Carmen Gregori-Signes and José Santaemilia Ruiz eds. *Multiperspectives in Analysis and Corpus Design*. Granada: Comares, 89–102.

Nickerson, Catherine and Elizabeth de Groot. 2005. Dear shareholder, dear stockholder, dear stakeholder: The business letter genre in the annual general report. In Paul Gillaerts and Maurizio Gotti eds. *Genre Variation in Business Letters*. Bern: Peter Lang, 325–246.

Patelli, Lorena and Matteo Pedrini. 2014. Is the optimism in CEO's letters to shareholders sincere? Impression management versus communicative action during the economic crisis. *Journal of Business Ethics* 124/1: 19–34.

Poole, Robert. 2016. Good times, bad times: A keyword analysis of letters to shareholders of two fortune 500 banking institutions. *International Journal of Business Communication* 53/1: 55–73.

Poole, Robert. 2017. "New opportunities" and "strong performance": Evaluative adjectives in letters to shareholders and potential for pedagogically-downsized specialized corpora. *English for Specific Purposes* 47: 40–51.

Real Academia Española. n.d. *Diccionario de la Lengua Española* (23rd ed., 23.5 online version). https://dle.rae.es/

Ruiz-Garrido, Miguel, Inmaculada Fortanet-Gómez and Juan Carlos Palmer-Silveira. 2012. Introducing British and Spanish companies to investors: Building the corporate image through the chairman's statement. In Jolanta Aritz and Robyn C. Walker eds. *Discourse Perspectives on Organizational Communication*. Plymouth: Fairleigh Dickinson University Press, 159–178.

Skorczynska Sznajder, Hanna. 2021. The analysis of adjectives in letters to shareholders of British and Spanish companies from the Valuation Theory. *Pragmalinguistics* 29: 421–439.

Skorczynska Sznajder, Hanna and Rosa Giménez-Moreno. 2016. Variation in letters to shareholders from British, Polish and Spanish companies: A comparative study. *Journal of Intercultural Communication* 40: 1–21.

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

Taylor, Charlotte. 2013. Searching for similarity using corpus-assisted discourse studies. *Corpora* 8/1: 81–113.

Taylor, Charlotte. 2018. Similarity. In Charlotte Taylor and Anna Marchi eds. *Corpus Approaches to Discourse: A Critical Review*. London: Routledge, 19–37.

Thompson, Geoff and Susan Hunston. 2000. Evaluation: An introduction. In Susan Hunston and Geoff Thompson eds. *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: Oxford University Press, 1–27.

Vendler, Zeno. 1957. Verbs and times. *The Philosophical Review* 66/2: 143–160.

Vogel, Radek. 2020. Persuasion in business discourse: Strategic use of evaluative lexical means in corporate annual reports. In Olga Dontcheva-Navratilova, Martin Adam, Renata Povolná and Radek Vogel eds. *Persuasion in Specialised Discourses*. Cham: Palgrave McMillan, 159–195.

Wang, Jinxiao. 2020. A corpus-based analysis of collocations in Chinese and American letters to shareholders. *International Journal Advances in Social Science and Humanities* 8/6: 10–19.

Corresponding author
Blanca Carbajo Coronado
Facultad de Filosofía y Letras
Autonomous University of Madrid
Francisco Tomás y Valiente, 1
28049 Madrid
Spain
E-mail: blanca.carbajo@uam.es

received: December 2022
accepted: March 2023

APPENDICES

Appendix 1: The 50 most frequent verbs in the profit and loss corpora

	Profit corpus	**Normalised frequency**	**Loss corpus**	**Normalised frequency**
1	*Ser*	9.43	*Ser*	10.21
2	*Estar*	3.19	*Estar*	3.85
3	*Tener*	2.58	*Tener*	3.18
4	*Permitir*	2.17	*Hacer*	2.20
5	*Seguir*	2.14	*Poder*	1.72
6	*Hacer*	1.89	*Permitir*	1.52
7	*Alcanzar*	1.69	*Seguir*	1.48
8	*Poder*	1.30	*Alcanzar*	1.47
9	*Mejorar*	1.27	*Pasar*	1.47
10	*Mantener*	1.14	*Querer*	1.13
11	*Crecer*	1.09	*Deber*	1.13
12	*Querer*	1.09	*Mantener*	1.04
13	*Realizar*	1.06	*Dar*	1.02
14	*Destacar*	1.06	*Continuar*	0.94
15	*Pasar*	1.03	*Llevar*	0.93
16	*Desarrollar*	1.02	*Trabajar*	0.88
17	*Unir*	1.00	*Suponer*	0.86
18	*Conseguir*	0.98	*Unir*	0.81
19	*Ir*	0.94	*Mejorar*	0.80
20	*Dar*	0.92	*Consolidar*	0.77
21	*Suponer*	0.89	*Reducir*	0.75
22	*Generar*	0.80	*Ver*	0.73
23	*Reducir*	0.77	*Poner*	0.73
24	*Deber*	0.70	*Lograr*	0.73
25	*Aumentar*	0.70	*Estimar*	0.73
26	*Trabajar*	0.70	*Conseguir*	0.73
27	*Avanzar*	0.69	*Crecer*	0.72
28	*Continuar*	0.69	*Desarrollar*	0.70
29	*Impulsar*	0.69	*Ir*	0.67
30	*Situar*	0.67	*Contar*	0.67
31	*Contar*	0.67	*Destacar*	0.65
32	*Poner*	0.66	*Realizar*	0.65
33	*Cumplir*	0.64	*Esperar*	0.64
34	*Llevar*	0.64	*Avanzar*	0.62
35	*Consolidar*	0.59	*Presentar*	0.61
36	*Superar*	0.58	*Decir*	0.59

37	*Esperar*	0.58	*Producir*	0.54
38	*Marcar*	0.56	*Obtener*	0.54
39	*Demostrar*	0.56	*Reforzar*	0.51
40	*Afrontar*	0.55	*Mostrar*	0.51
41	*Estimar*	0.53	*Afrontar*	0.48
42	*Contribuir*	0.50	*Centrar*	0.48
43	*Obtener*	0.50	*Iniciar*	0.48
44	*Ver*	0.48	*Cerrar*	0.48
45	*Prever*	0.47	*Generar*	0.48
46	*Presentar*	0.47	*Situar*	0.45
47	*Ascender*	0.47	*Registrar*	0.45
48	*Producir*	0.47	*Reflejar*	0.41
49	*Operar*	0.45	*Marcar*	0.41
50	*Agradecer*	0.44	*Llegar*	0.41

Appendix 2: The 50 most frequent eventive nouns in the profit and loss corpora

	Profit corpus	Normalised frequency	Loss corpus	Normalised frequency
1	*Ejercicio*	3.17	*Negocio*	2.98
2	*Resultado*	2.91	*Ejercicio*	2.63
3	*Crecimiento*	2.77	*Resultado*	2.20
4	*Negocio*	2.42	*Crecimiento*	1.61
5	*Compromiso*	1.36	*Venta*	1.53
6	*Mejora*	1.31	*Proceso*	1.31
7	*Venta*	1.27	*Gestión*	1.28
8	*Desarrollo*	1.17	*Cambio*	1.13
9	*Esfuerzo*	1.17	*Deuda*	1.13
10	*Producción*	1.14	*Coste*	0.97
11	*Inversión*	1.09	*Desarrollo*	0.96
12	*Gestión*	1.02	*Reducción*	0.88
13	*Administración*	0.84	*Compromiso*	0.85
14	*Evolución*	0.83	*Esfuerzo*	0.85
15	*Proceso*	0.83	*Administración*	0.83
16	*Acuerdo*	0.77	*Mejora*	0.83
17	*Innovación*	0.72	*Construcción*	0.75
18	*Reducción*	0.70	*Onversión*	0.73
19	*Coste*	0.69	*Cierre*	0.72
20	*Trabajo*	0.67	*Operación*	0.70
21	*Consumo*	0.64	*Trabajo*	0.70
22	*Cambio*	0.63	*Acuerdo*	0.67
23	*Gobierno*	0.63	*Generación*	0.64
24	*Ingreso*	0.56	*Transformación*	0.64
25	*Beneficio*	0.54	*Ampliación*	0.61
26	*Deuda*	0.54	*Evolución*	0.61
27	*Comportamiento*	0.52	*Ingreso*	0.57
28	*Construcción*	0.52	*Gasto*	0.51
29	*Emisión*	0.50	*Pérdida*	0.49
30	*Generación*	0.48	*Producción*	0.49
31	*Operación*	0.48	*Reestructuración*	0.49
32	*Recuperación*	0.48	*Integración*	0.41
33	*Cierre*	0.47	*Fobierno*	0.40
34	*Gasto*	0.43	*Aumento*	0.38
35	*Aumento*	0.42	*Consumo*	0.38
36	*Cumplimiento*	0.42	*Ahorro*	0.37
37	*Iniciativa*	0.42	*Caída*	0.37
38	*Transformación*	0.39	*Incremento*	0.37

39	*Solución*	0.38	*Innovación*	0.35
40	*Adquisición*	0.36	*Organización*	0.35
41	*Creación*	0.36	*Actuación*	0.34
42	*Demanda*	0.36	*Beneficio*	0.34
43	*Incremento*	0.36	*Comportamiento*	0.34
44	*Actuación*	0.30	*Compra*	0.34
45	*Instalación*	0.30	*Control*	0.34
46	*Aprobación*	0.28	*Oferta*	0.34
47	*Control*	0.28	*Iniciativa*	0.32
48	*División*	0.28	*Solución*	0.32
49	*Pago*	0.28	*Comercialización*	0.30
50	*Caída*	0.27	*Demanda*	0.30

RiCL Research in Corpus Linguistics

Multicultural London English (MLE) as perceived by the press, on social media, and speakers themselves

Ignacio M. Palacios-Martínez
University of Santiago de Compostela / Spain

Abstract – This paper aims to contribute to the study of Multicultural London English (MLE) by focusing on the perceptions of MLE speakers of their own linguistic production and, also, by exploring the reactions and responses to this variety in the British press and on social media. The results indicate that most of the MLE speakers feel that they use a kind of slang. The majority of accounts found in the media depict MLE as foreign, associated with grime music and bad behaviour. Opinions garnered from social networks show more diverse views; while some reiterate the perceived negative aspects, others highlight its multicultural nature and uniqueness. The paper also suggests measures that could be adopted to change negative attitudes towards MLE.

Keywords – Multicultural London English; language attitudes; Cockney; teenagers' language; language contact; multiethnolect

1. INTRODUCTION[1]

Over the last two decades a new multiethnolect (Clyne 2000)[2] has emerged in London, widely known as Multicultural London English (henceforth, MLE) —see Cheshire *et al.* (2011) or Cheshire (2019)— but also as New Cockney (Fox 2015) or even as Jafaican/Jafaikan, that is, fake Jamaican,[3] because it is generally believed that a large

[1] I sincerely thank the editors and the two reviewers for taking the time to review the manuscript and providing constructive feedback to improve the original. For generous financial support, I am grateful to the following institutions: The Spanish Ministry of Science and Innovation (grant PID2021-122267NB-00), the European Regional Development Fund (grant PID2021-122267NB-00), and the Regional Government of Galicia (Consellería de Educación, Cultura e Universidade, grant ED431B 2021/02).

[2] A multiethnolect is, according to Clyne (2000: 87), an ethnolect where members of the dominant group, particularly young speakers, share it with other ethnic minorities in a language-crossing situation. This is regarded as "the expression of a new kind of group identity." The concept has also been referred to as 'contemporary urban vernaculars' (Rampton 2015), 'urban vernacular' and 'urban youth speech style' (Wiese 2009; Cheshire *et al.* 2015; Nortier and Svendsen 2015) and even, more recently, as 'urban contact dialect' (Kerswill and Wiese 2022).

[3] Kerswill and Torgersen (2021) show how the influence of Jamaican English in MLE is particularly visible at the lexical level but not so much in the morphosyntax (except for the pronoun *man*, which is usually equivalent to the first or third singular personal pronouns in English), and in phonology.

number of its speakers use an accent and expressions typical of the Caribbean, more particularly from Jamaica. However, MLE is much more than this, in that it has been formed by a feature pool (Mufwene 2001) derived from local varieties (namely Cockney), plus other UK dialects of English, standardised varieties of English, in addition to the expression of an array of speakers from different Caribbean, Indian, North-African and Asian backgrounds. Similar developments have taken place in other multilingual European and African cities (Wiese 2009; Kerswill and Wiese 2022) and even within the UK, to the extent that some scholars such as Drummond (2018) refer to the existence of a Multicultural Urban British English.

There is a growing literature on many of its innovative phonetic, lexical, grammatical and discourse features, including quotatives (Fox 2012), intensifiers (Núñez and Palacios-Martínez 2018), pragmatic markers (Palacios-Martínez 2015; Torgersen *et al.* 2018), negatives (Lucas and Willis 2012; Palacios-Martínez 2016, 2017), address terms (Palacios-Martínez 2018), verb variation (Cheshire and Fox 2009) and how certain of its vowels and consonants have a different pronunciation from standardised varieties of English (Cheshire *et al.* 2011; Fox 2015). To these investigations, we might add studies focusing on the attitudes of both MLE and non-MLE speakers towards the sociolect[4] itself (Kerswill 2013, 2014; Cardoso *et al.* 2019; Gates and Ilbury 2019; Kircher and Fox 2019a, 2019b; Levon *et al.* 2021; Sharma *et al.* 2022).

The current study seeks to contribute to this body of work by investigating the perceptions of MLE speakers towards their own variety, that is, their attitudes and perceptions of the language they use in their everyday lives, and also by considering the reactions and responses towards MLE in a variety of British media and on social networks. To this end, the analysis will be based on materials extracted from the *London English Corpus* (LEC; Cheshire *et al.* 2011), newspapers, radio and TV programmes, together with posts from *Twitter* and videos available on *YouTube*, along with their corresponding comments.

The paper is organised as follows. Following this introduction, the concept of language attitudes to be used in this study will be defined in Section 2, noting the different approaches taken in research, and justifying those to be used here. This will be followed

[4] The terms ‘sociolect’, ‘ethnolect’ and ‘multiethnolect’ can be used interchangeably since they basically express the same meaning. However, ‘sociolect’ is a more neutral label, while ‘ethnolect’ refers to a language variety associated with a particular ethnic group. A ‘multiethnolect’ is, in fact, a type of ethnolect, as stated in footnote 2.

by a review of existing studies on the description of language attitudes and ideologies in MLE (Section 3). Section 4 will deal with the objectives and methodology of the study and will provide a section setting out the main findings. This latter will be organised around three major headings: 1) speakers' perceptions of their own mode of expression, 2) MLE as perceived in the media and on *Twitter*, and 3) the presence of MLE on *YouTube*, together with viewer comments and reactions. Following this, Section 6 will be concerned with a description and some reflections on certain measures that could be taken to engender positive views on MLE and its speakers in educational settings. This is an important issue, in that attitudes of acceptance and tolerance towards non-standard or non-mainstream varieties and their respective speakers should be fostered by educational authorities, social institutions and in the mass media. The paper will conclude with a summary of the main findings in Section 7.

2. Defining and investigating language attitudes

Language attitudes (henceforth, LA) are the ideas and opinions, beliefs and prejudices that speakers hold towards a particular language, variety or accent as a whole, or towards a specific feature of any of these (Oppenheim 1982: 39). However, the field of LA is not limited to this and is, in fact, rather broad. For example, Baker (1992: 29) refers to a number of domains within the scope of LA which cover areas such as attitudes to language variation, the learning of a new language, attitudes to particular language lessons, language preferences and parents' views on language learning.

LA can also be seen as the study of reactions or responses to a particular stimulus, which —in this case— might simply be exposure to the variety in question. Three main dimensions or components can be distinguished (Garrett 2010: 23): 1) a cognitive element, which corresponds to a speaker's beliefs and opinions; 2) an affective dimension, having to do with feelings and emotions; and 3) a conative constituent, responsible for our behaviour, reactions and responses. This is generally known in the literature as the ABC model of attitudes, which is based on Baker (1992) and Augoustinos *et al.* (2006), although the latter studies go back ultimately to Rosenberg and Hovland (1960) and Azjen (1988). Furthermore, according to Garrett (2010) and Dragojevic *et al.* (2017), LA are organised along two evaluative dimensions which are present in the majority of LA studies: 1) status (e.g., intelligent, educated, competent) and 2) solidarity (e.g., friendly, unpleasant). Status attributions refer mainly to the individual's perceptions

of socioeconomic conditions, while solidarity tends to be based on in-group loyalty, that is, the degree to which the speaker is perceived as being a close or distant member of the group. In this study, this would apply to the perceptions of the speakers' status in MLE and also to their degree of identification with speakers who represent this variety.

There has been considerable debate as to the origins of these LA. Although for quite some time it was argued that they are mainly innate, it is now widely thought that they are also learned, that is, we tend to be influenced by the attitudes of society as a whole and the people around us (Allport 1954). As Oppenheim (1982: 40) claims, "they are more likely to have been adopted or taken over from significant others as part of our culture and socialization." From an early age, children develop an awareness of the language they use and tend to show a preference for their own language variety (Ebner 2007: 64). In our present data, children as young as eight are able to distinguish the accent used in one London area from that typical of another neighbourhood in their community, and they are also able to discuss these. However, this does not mean that underlying attitudes towards varieties cannot be changed. Language attitudes commonly come hand in hand with stereotypes, ones which are often not justified. In this respect, standard varieties tend to be associated with prestigious, well-educated and middle/high class individuals, while non-standard ones are often seen as rude, uninformed and typical of ignorant working-class or lower-class members of society (Trudgill 1975; Milroy 2001). This is a relevant issue to be taken into consideration in the educational field since the sort of information children receive and the type of attitudes fostered by teachers and educators with respect to the status and role of the languages studied, or even towards their own variety, will be of vital importance. As Trudgill (1975: 61) rightly claims:

> teachers' attitudes to children's language can be very influential in shaping relationships between the child and the school, and in affecting a child's attitude to education generally.

In the study of language attitudes three main kinds of research methods can be distinguished (Garrett *et al.* 2003: 14–18; Garrett 2010: 37–52; Kircher and Zipp 2022): the societal-treatment approach, direct approaches and indirect ones. In the first of these, researchers gather attitudes from observed behaviours, and subsequent analysis focuses on the treatment of language and language varieties, the study of government and educational language-policy documents views on the use of various languages in education, the use of dialect forms in the literature, the discourse analysis of print media and content analysis of social media (including social networks and other digital genres).

In turn, so-called direct approaches are based on the elicitation of data. Informants are asked to report their attitudes through scales, questionnaires, surveys, polls, interviews, focus groups or through the methodologies of perceptual dialectology. Corpora studies (Vessey 2015) might also be classified within this group. Finally, indirect approaches involve techniques that go beyond asking direct questions, and often adopt the Matched Guise Technique (MGT) developed in the late 1950s by Lambert and his colleagues in Canada.[5]

3. PREVIOUS RESEARCH

The field of LA has been investigated extensively from both psychological and sociolinguistic perspectives, and this also seems to be the case when considering LA as applied to MLE, especially taking into account that it is quite a new dialect. I will focus on these studies in the following section.

Kerswill (2013) deals with the construction of language by young speakers of London considering their beliefs and views on the issues of identity, place and ethnicity. For this purpose, he studies the speakers' own perceptions and constructions of speech produced in inner and outer London, Hackney and Havering. The results indicate that, as regards Hackney speakers, those who are 'not Anglo' do not identify themselves with Cockney, either as a group identity or as their mode of expression, while the opposite applies to a small group of 'Anglo' speakers. Likewise, a similar number of both groups consider that they use a kind of slang. This clearly contrasts with the views of Havering speakers, who associate themselves with slang and do not claim a Cockney identity. Although these findings are quite revealing, the distinction made between 'Anglo' and 'non-Anglo' is somewhat blurred and can be easily questioned nowadays.

Kerswill (2014) also considers the presence of MLE in the media, specifically in reactions in the British press between 2000 and 2013. His analysis of comments therein illustrates that MLE is regarded as a threat, and that there are two essential components to this. The first of these has to do with the displacement of Cockney and the loss of British cultural values, while the second is a threat to liberal principles (gender equality

[5] In MGT, interviewees are asked to respond according to different criteria to the varieties of speakers whose voices are recorded on tape, whereby the same speaker uses different linguistic varieties or accents, something the interviewees are not generally well aware of. This accounts for the label given to this technique: 'matched guise'.

and homosexual equality). Likewise, MLE is associated with bad behaviour and, more particularly, with the social unrest and riots that took place in London in August 2011.

Gates and Ilbury's (2019) paper is broader in scope and considers how standard ideologies can constrain and affect speakers of non-standardised varieties. To this end, they analyse data collected from two groups of young speakers from different areas of London between 2015 and 2017. MLE is characterised by the young participants in this study as being 'urban' and 'street-ready', in contrast to the 'standard', and is regarded as inappropriate for the classroom. In addition, adolescents are aware of certain stigmas associated with some vernacular forms they use, and a connection between language and race is also drawn since, in the views of participants in the study, white speakers tend to speak more formally than black ones. Young learners are also aware of the importance of using standardised varieties of English. Apart from this, the authors maintain that some tension between the curriculum and the way people use language everyday is identified. Thus, according to the answers given by the participants in the study, those forms of speaking which are regarded as more formal (i.e., standardised varieties of English) are considered important for the future, for education and for getting a job, but not for every day social interactions. In contrast, any way of speaking that does not follow the standard is not perceived as 'normal', that is, as following the mainstream, and a stigma is seen to be attached to it.

Kircher and Fox (2019a, 2019b) focus specifically on attitudes towards MLE, whilst also investigating the implications of these for attitude theory and language planning. Findings indicate that the classic status-solidarity distinction is not confirmed, this being regarded as unusual. The authors argue that this may have been related to the fact that MLE does not behave like other language varieties and has its own characteristics as a multiethnolect. The participants' overall attitudes towards MLE were negative, although speakers of MLE held more positive attitudes towards their own variety. Among the factors that had an impact on the creation of LA was the contact with MLE speakers which fostered positive opinions. Speakers of languages other than English maintained more positive views towards MLE and the same applied to speakers with high levels of education. Kircher and Fox (2019a, 2019b) conclude by noting the need to engender positive attitudes towards MLE and its speakers through the reduction of stereotypes.

Kircher and Fox (2019b) addresses the issue of standard language ideologies in relation to MLE. The data were collected through an online questionnaire, conducted

between October 2016 and July 2017. Regarding language ideologies, the data reveal that non-MLE-speaking Londoners used more negative than positive terms to describe the multiethnolect, while MLE speakers themselves resorted to more positive labels. The negative semantic categories used to describe MLE include terms such as broken language, language decay, secret code and fake variety. In contrast, positive semantic categories describe MLE as mainstream, a natural evolution, cool, interesting, fascinating, innovative, endearing, rich and relaxed.

As far as social stereotypes are concerned, the data analysed in Kircher and Fox (2019b) contain numerous representations of MLE speakers' demographic characteristics: ethnic minorities, age (teenagers), class (working-class), gender (male users) and location (East End of London), with non-MLE speakers maintaining stronger social stereotypes here. The negative stereotypical characteristics attached to MLE speakers were those of aggression, lack of education and intelligence, and the inability to switch to the standard language.

Cardoso *et al.* (2019) describe and illustrate the importance of inter-speaker variation in the evaluation of British accents as part of a nationwide survey based on interviews conducted with a sample of 1,015 participants. In their analysis of five British accents, special attention is paid to MLE. Speakers with standard accents, such as RP, are more positively rated that those showing a southern accent such as Estuary English or MLE, the latter being the lowest rated of all. Non-standard northern accents, in turn, stand between these two poles. The authors also conclude that those speakers of MLE with more accentuated MLE features, such as *k*-backing or *th*-stopping, trigger more negative attitudes, since these accent traits tend to be associated with specific socio-indexical traits (being less educated and ethnically black). Moreover, accent bias seems to be present to some degree in employment contexts, to the extent that MLE speakers with a more clearly distinctive accent are more negatively evaluated in terms of hireability.

Also, Levon *et al.* (2021) report on a large-scale study focusing on current attitudes to accents in England. Through a verbal guise technique, a sample of 848 raters evaluated the interview performance and potential hireability of candidates for a position in a law firm. These candidates were native speakers of English who showed one of the five characteristic accents of England (RP, Estuary English, MLE, General Northern English and Urban West Yorkshire English) in their speech. Results indicate that bias persists in British society against particular accents such as Estuary English and MLE. The authors

also examine the impact that this may have in perpetuating social inequalities in England with the implications that this has in the labour and educational fields.

Finally, Sharma *et al.* (2022) present an updated overview of national attitudes towards various accents by replicating and expanding previous studies. In this study, a total of 821 British subjects, with age ranging from 18 to 79, were asked to rate 38 accents on a seven-point scale for prestige and pleasantness. The results show that some conservative accents are demoted in terms of perceived prestige, while some other lower-ranking ones are more positively considered than was previously the case. MLE itself is found to be in nineteenth position regarding prestige, with an average rating of 3.81, whereas it occupies twenty-fourth position when rated for pleasantness. RP, and so-called Queen's English, plus French accents are the most favourably rated in both categories. Furthermore, the authors also conclude that the hierarchy of accent prestige is conditioned by a number of social, contextual and psychological factors, such as the respondent's age and regional origin, together with stimulus content and a respondent's psychological predisposition.

4. Purpose and method

The purpose of the present study is to deepen our understanding of attitudes towards MLE by investigating how its speakers see themselves, and how they are perceived in the broadcast and traditional printed media, in social networks, and thus in wide sectors of society generally. As with previous studies, I use recent data drawn from corpora, newspapers, radio programmes and social media such as *Twitter* and *YouTube*, employing a combined approach to the study of attitudes towards MLE. Given the rise of digital genres as forms of communication over the last two decades (Squires 2016; Herring 2019) and their attested value as useful sources for language research (Palacios-Martínez 2020), together with the growing importance of social networks for the young and middle-aged generations, an analysis of data from these sources may help us to gain a better understanding of the attitudes towards MLE, the representation of this variety in the media and on some social networks, and the possible implications for language planning and education. In this respect, I also intend to reflect on possible measures to change the negative attitudes towards MLE identified in this and previous studies. All this aims to contribute to the understanding of attitudes to MLE and to its perception in the media and

on social networks, which thus far has adopted a direct approach by considering data largely from questionnaires and interviews.

The method followed here can be defined as mixed, combining the direct and societal treatment approaches described in Section 2. For the direct element, I will use data from LEC, compiled by Cheshire and her team in London between 2004 and 2010 (Cheshire *et al.* 2011; Cheshire 2019), which consists of the *Linguistic Innovators Corpus* (LIC)[6] and the *Multicultural London English Corpus* (MLEC).[7] The data for the former corpus, which contains over a million words from 121 speakers, was collected between 2004 and 2007 in the districts of Hackney (inner London) and Havering (outer London) and includes the speech of both teenagers and adults. The MLEC was compiled between 2007 and 2010 and contains data not only from young speakers but also from children as well as from different adult speaker groups, covering parts of the districts of Islington, Haringey and Hackney in north London. It amounts to 621,327 words from a total of 137 speakers. In both cases, the material was collected through individual and group interviews in youth centres and schools.

The LEC corpus was accessed using *SketchEngine* (Kilgarriff *et al*. 2014), which allowed me to conduct different types of simple and combined queries. For the extraction of the data the following key words directly connected with language and related terms were searched: *Cockney*, *language*, *speech*, *talk*, *jargon*, *lingo*, *slang*, *accent*, *London English* and *standard*. Once all the tokens were retrieved, they were manually analysed in accordance with the purposes of the study. In addition, I reviewed all those newspaper articles that mentioned MLE and/or London English from 2011 to 2020. This particular period was selected because Kerswill (2014) had already surveyed the timespan between 2000 and 2013, and hence it was of interest to see what had happened between 2013 and the present. To this end, I followed a procedure similar to that used by Kerswill (2014) by searching *Nexis UK*,[8] an online database of English language newspapers and other media known, for all contributions referring to the English language; labels here included *Cockney*, *Jafaican/Jafaikan*, *London English*, *London accent* and *MLE*. News and other articles from seven daily papers were retrieved and examined closely together with BBC reports, both on TV and radio, for the same period. I then turned to attitudes, perceptions

[6] https://www.lancaster.ac.uk/fss/projects/linguistics/innovators/
[7] https://www.sketchengine.eu/london-english-corpus/
[8] https://bis.lexisnexis.co.uk/research-and-insights/nexis

and reactions towards MLE in social networks, starting with *Twitter* in general and then focusing on exchanges of three rappers who are frequently identified with MLE, namely Dizzee Rascal, Wiley and Dappy (the three of them stage names). The accounts of these three artists were selected because they were brought up in London, the first two specifically in East London, an area that has been traditionally associated with the origin of Cockney; all three have a significant impact on the music industry, and they all make overt, public use of this sociolect in their everyday communication and in their exchanges with their fans and followers. The analysis of the *Twitter* material was restricted to the last 15 years and included not only the tweets posted by the three rappers in question but also all the responses and reactions of their fans and followers. It must be borne in mind that the responses given by the speakers vary greatly in terms of their length and the kinds of details provided, with some of the respondents providing very elaborate answers, while others being more sparing with words. The previous data were complemented by the examination of videos about MLE and London English available on *YouTube*, looking not only at their content but also at the comments below a video, which yielded valuable information regarding the views and opinions of individual users. The analysis of all this data will be mainly qualitative, although some figures will also be provided to better illustrate some of the points made. This study is thus intended to make a contribution to previous research by providing the perspective of speakers in media and social networks together with that of the MLE speakers themselves. The information and recent data obtained from the press and social networks will hopefully serve to complement the findings of previous studies.

5. FINDINGS

5.1. Speakers' perception of their own variety

I here focus specifically on the speakers' definition and description of the type of language or expression the participants of the different age groups (adolescent, teenagers, young adults, middle-aged adults and elderly speakers) think they use rather than in terms of their ethnicity and identity. As mentioned above, the latter was closely analysed by Kerswill (2013) although, in his account, he was restricted only to the language of the young speakers, while now new data extracted from a longer and more recent period of time and from the rest of the age groups are also considered.

As noted above, data from LEC were the main source used to investigate this issue. The word *Cockney(s)*, referring either to the language variety or its community of speakers, occurs 244 times. This high number of tokens in LEC stems from the questions the fieldworkers ask participants about the accent or the type of language they think they use, and also whether they identify themselves with Cockney or not. Apart from this, there are also a high number of repetitions typical of spoken language.

A close look at the data shows that 28.3 per cent of the respondents identify themselves with Cockney, 41.7 per cent claim that they use some kind of slang while 5.9 per cent opt for patois. Other terms they mention to designate their mode of expression are the following: *gangsta, east London Cockney*, *urban speech* (*bashment*), *street talk*, *new lingo*, *Hackney Cockney* and *London accent*. The area of London where they live and even at times their ethnicity may have a bearing on their decision. Thus, the majority of the respondents who choose the label *Cockney* come from inner London and are white and Anglo speakers, while those who select *patois* are of African Caribbean origin. As regards the term *slang*, views are more divided according to the area of London participants come from, although, in this case, it is the clearly the preferred alternative for non-Anglos. Some examples are provided in (1)–(4).

(1) William (17 years, inner London): What we call it is *urban speech*.

(2) Mandy (16 years, outer London, Havering): We are tipical cockneyes the way we talk and that we talk in *slang*.

(3) Robert (16 years, inner London): We call it *urban speech gansta*.

(4) Alan (age unknown) yea just *street talk* it's like … slang. It's all sort slang when we talk.

Table 1, below, sets out the different terms used by respondents according to their age group to refer to their own expression.

COCKNEY					
Speaker's age	**No.**	**London Area**		**Ethnicity**	
12	**2**	Inner London	2	Anglo	1
		Havering (Outer London)	-	Non-Anglo	1
16–19	**12**	Inner London	8	Anglo	8
		Havering (Outer London)	4	Non-Anglo	4
20–30	**1**	Inner London	1	Anglo	1
		Havering (Outer London)	-	Non-Anglo	-
40–50	**-**	Inner London	-	Anglo	-
		Havering (Outer London)	-	Non-Anglo	-
+70	**4**	Inner London	3	Anglo	4
		Havering (Outer London)	1	Non-Anglo	-
TOTAL	**19**				

SLANG (ING)					
Speaker's age	**No.**	**London Area**		**Ethnicity**	
12	**3**	Inner London	3	Anglo	1
		Havering (Outer London)	-	Non-Anglo	2
16–19	**21**	Inner London	11	Anglo	6
		Havering (Outer London)	10	Non-Anglo	15
20–30	**-**	Inner London	-	Anglo	-
		Havering (Outer London)	-	Non-Anglo	-
40–50	**4**	Inner London	-	Anglo	2
		Havering (Outer London)	4	Non-Anglo	2
+70	**-**	Inner London	-	Anglo	-
		Havering (Outer London)	-	Non-Anglo	-
TOTAL	**28**				

PATOIS					
Speaker's age	**No.**	**London Area**		**Ethnicity**	
16–19	**3**	Inner London	3	Anglo	-
		Havering (Outer London)	-	Non-Anglo	3
40–50	**1**	Inner London	1	Anglo	-
		Havering (Outer London)	-	Non-Anglo	1
TOTAL	**4**				

Table 1: Terms used by the respondents in LEC to describe the kind of language they use according to age group, London area and ethnicity

NORMAL / COMMON					
Speaker's age	**No.**	**London Area**		**Ethnicity**	
12	**2**	Inner London	2	Anglo	1
		Havering (Outer London)	-	Non-Anglo	1
16–19	**1**	Inner London	1	Anglo	-
		Havering (Outer London)	-	Non-Anglo	1
TOTAL	**3**				

GANSTA					
Speaker's age	**No.**	**London Area**		**Ethnicity**	
16–19	**3**	Inner London	2	Anglo	-
		Havering (Outer London)	1	Non-Anglo	3
TOTAL	**3**				

EAST LONDON COCKNEY					
Speaker's age	**No.**	**London Area**		**Ethnicity**	
16–19	**2**	Inner London	2	Anglo	1
		Havering (Outer London)	-	Non-Anglo	1
TOTAL	**2**				

STREET TALK					
Speaker's age	**No.**	**London Area**		**Ethnicity**	
16–19	**1**	Inner London	-	Anglo	-
		Havering (Outer London)	1	Non-Anglo	1
TOTAL	**1**				

DIALECT					
Speaker's age	**No.**	**London Area**		**Ethnicity**	
+70	**1**	Inner London	1	Anglo	1
		Havering (Outer London)	-	Non-Anglo	-
TOTAL	**1**				

Table 1: (Continuation)

URBAN SPEECH					
Speaker's age	**No.**	**London Area**		**Ethnicity**	
16–19	**1**	Inner London	1	Anglo	-
		Havering (Outer London)	-	Non-Anglo	1
TOTAL	**1**				

COCKNEY SLANG					
Speaker's age	**No.**	**London Area**		**Ethnicity**	
16–19	**1**	Inner London	-	Anglo	1
		Havering (Outer London)	1	Non-Anglo	-
TOTAL	**1**				

HACKNEY COCKNEY					
Speaker's age	**No.**	**London Area**		**Ethnicity**	
16–19	**1**	Inner London	1	Anglo	1
		Havering (Outer London)	-	Non-Anglo	-
TOTAL	**1**				

HACKNEY STYLE GHETTO					
Speaker's age	**No.**	**London Area**		**Ethnicity**	
16–19	**1**	Inner London	1	Anglo	-
		Havering (Outer London)	-	Non-Anglo	1
TOTAL	**1**				

DIFFERENT LINGO					
Speaker's age	**No.**	**London Area**		**Ethnicity**	
16–19	**1**	Inner London	1	Anglo	1
		Havering (Outer London)	-	Non-Anglo	-
TOTAL	**1**				

NEW LINGO					
Speaker's age	**No.**	**London Area**		**Ethnicity**	
16–19	**1**	Inner London	1	Anglo	-
		Havering (Outer London)	-	Non-Anglo	1
TOTAL	**1**				

Table 1 (continuation)

When interpreting these data, we ought to bear in mind that the number of speakers for each age group is not the same, the 16–19 year group being the largest in number; also, for some speakers these labels are not mutually exclusive. Hence, it is relatively common that they use two or three of these labels, claiming as they do that they can adapt and switch their expression according to the situation or interlocutor in question, or even according to the communicative purpose intended so as to sound funny or make fun of someone. This seems to be particularly frequent in the case of middle-aged and elderly speakers, as shown in (5)–(6).

(5) Talulah's father (45 years, inner London): when we'd meet like . of s say for instance I'd meet I'd I a white person or yeah no I talk to my brother say my brother [right okay] . and I would say . whagwan uhu .. theirs is much .. you can hear their English .. they the . the the broken up . English . Jamaican patois ... and it would sound it would sound totally different.

(6) Serena (18 years, inner London): sometimes I'll be in a cockney mode sometime. I'll be in like a ghetto mode.

A large number of these participants do not know how to classify themselves (cf. (7)), that is, they cannot think of a specific name for their variety or accent, and none of the speakers uses the label *MLE* or *Jafaican*.

(7) Interviewer: How would you describe yourself
Justin (16 years, outer London). I'm not like cockney or nothing like my family. I'm just common but erm I dunno.

As regards the association of Cockney with a particular area or neighbourhood of London, there are also some elderly speakers who associate Cockney particularly with the East End of London (cf. (8)).

(8) Joe (70+ years, inner London): People say you are cockney but a cockney is strictly within the sound of Bow bells mm supposed to be yeah supposed to be.

However, there are no unanimous views on this since for some other respondents there are now more speakers of Cockney outside London (in Essex, for example) than in the capital itself, something that has also been pointed out by Fox (2015: 29), due to population movements and the arrival of immigrants (cf. (9)):

> The white working-class families- the 'Cockneys'- have, in the main, left the area and moved out to the suburbs of London, Essex and surrounding areas. In doing so, it might be said that they have caused the geographical 'spread' of the East End, this term now being applied to a much wider area than that with which it was traditionally associated.

(9) Ted (+70 years, inner London): most of the east like east enders cockneys moved out to essex and they're cockney lang. (LEC)

The age factor seems to play a role here, in that some of the respondents make a distinction between the type of language used by teenagers and that typical of adults: for some speakers, Cockney is associated with the older generation, the *sweet people*, whereas the new form of speaking is connected with a younger age group, that is, the *safe people*, since the latter tend to use this expression very often in their everyday activities (cf. (10)).

(10) William (17 years, inner London): "Sweet people speak cockney, safe people use urban speech."

Attitudes to Cockney in particular vary greatly from one speaker to another. Thus, some of them maintain that Cockney is rude and a lazy way of speaking, as shown in (11).

(11) Ted (+70 years, inner London): i was lazy i suppose i was cockney . in a lot . cos cockney is a lazy way of speaking.

However, for some others it is a form of expression they all share, and they even refer to particular features of Cockney which they like and feel proud of because it makes them feel part of their own culture; this is the case with the accent, rhyming slang, and the use of the address terms *mate* and *geezer*, as in (12).

(12) Paul (16 years, inner London): "you alright mate" like everyone's using it so I I kind of like it you know # laughter # I won't even lie. I actually like it like the cockney accent's kind of big so . everyone using the cockney accent and mate at the end and . like "you alright you alright geezer" and all that.

For some of the respondents, Cockney is also associated with brusque speech, in contrast to standardised varieties of English, which sound softer in tone. It is also contrasted with 'posh' English and is considered to be fake. Thus, Cockney speakers are even regarded as performers by an elderly speaker (cf. (13)).

(13) Ted (+70 years, inner London): I've noticed that most cockneys are performers er. I noticed it most when I went into the army.

We can also find discussions on the issue of race and its possible connections with ways of speaking. While for some respondents the variety used by a speaker is conditioned by their race, for others the place or area where a speaker lives plays a far more significant role, as exemplified in (14):

(14) Sulema (18 years, inner London): I don't think white people black people speak differently it's just in the area which you're in .. that makes you you know cos if you see white people and black people in Hackney they all speak the same to me but then again if you go to . somewhere like . Chelsea side they will speak differently from how we speak here.

Some of the respondents also feel that the variety they use in their everyday communication would not be the one expected to be used in school, since it is clearly different from what they regard as standardised varieties of English.

It is also interesting to see that teenagers in particular are able, in their explanations, to identify and discuss the meaning and implications of certain words which are typical of their own mode of expression, namely ethnical and slang terms, such as *gash* for girl and *waste/road man*, to refer to someone who spends a lot of time on the streets, *creps* or *kreps* for trainers, *low batties* for low trousers, the exclamative *Oh my days* equivalent to *Oh my god!*, *bredren* and *bruvs* to refer to their peers, *geezer* for man, *nang* for cool, *sket*, a pejorative term to refer to a girl, *bait* as obvious or well-known, *chav* referring to a white working class person with a stereotyped lifestyle and way of dressing, and *ends* and *yard* for local area, etc. Some examples are provided in (15)–(17).

(15) Maria (18 years, inner London): think it's a actually a jamaican words i really do believe that they call trainers *kreps* [aah] in Jamaica.

(16) Maria (18 years, inner London): everyone's using it though *oh my days oh my days* . oh my god oh my god.

(17) Dale (17 years, outer London): low batties was invented by . blacks .. because of prison … well in prison they only had small medium and large sizes like for the trousers and tops and that

All these exchanges show that the participants are not only aware of the kind of language and accent they use but also possess some metalinguistic knowledge as to a number of its main features: slang words and expressions, degree of formality and level of acceptance by society and their teachers at school, questions related to identity, social class and race, etc. This does not apply only to middle-aged and elderly speakers but also to young speakers.

5.2. MLE in the media

It is important to explore how MLE is portrayed in the print and broadcast media to identify those features which seem to be the most relevant and attractive, and to confirm

the extent to which their descriptions and the information provided are accurate. It is also important to see how all this contrasts with the perceptions of the speakers themselves.

The current analysis covers a total of 17 articles and radio programmes dealing directly with MLE from February 2011 to November 2019. As noted above (Section 4), Kerswill (2014) already dealt with the period between 2000 and 2013. The year 2014 yielded no information, whereas in 2016 four articles appeared. Table 2 provides full details of the journal of publication, date, headlines, and main contents.

Source	Date	Headline	Main contents
The Evening Standard	01/02/11	Language can't stay still - just listen to London.	Cockney is losing ground and it may disappear in 30 years being supplanted by MLE.
	31/01/13	English still stands tall in multicultural London.	Teenagers who have never been in contact with Caribbean speakers introduce in their conversation words of Jamaican patois.
The Daily Star	14/03/11	Anuvahood 15.	Taking the series Anuvahood as the source of examples, the author maintains MLE speakers can be regarded as performers since they tend to portray a Jamaican accent.
National Association for the Teaching of English (NATE) Classroom Vol. 17.	22/06/12	A multicultural English language.	The perceived Jamaican influence on teenagers' speech is regarded as a problem in education.
The Daily Mail	25/07/13	Present Day Cockney Speakers more likely to live in Essex than the East End of London.	Cockney is giving space to MLE mainly because of immigration.
	11/10/13	Why are so many middle-class children speaking in Jamaican patois?	MLE is considered to be a kind of superbug infecting children, this having serious consequences for education and the job market
Mail online	10/11/13	Is this the end of Cockney? Hybrid dialect dubbed 'Multicultural London English' sweeps across the country.	Cockney is being replaced by MLE and is also spreading to other parts of England, such as Manchester and Birmingham.

Table 2: Overview of the attitudes towards MLE in the media examined

Source	Date	Headline	Main contents
Metro	25/09/15	My London... Dizraeli; The rapper and musician loves to escape to Waktthamstow … and is fascinated by London lingo.	An interview with this musician who claims he loves MLE because it is "crazy and rich." London kids are seen as living representations of modern times and teenagers are agents of language change and innovation.
The Independent	05/01/16	Youth slang decoded: How to tell a 'durkboi' from a 'wasteman', bruv.	In defence of youth language and slang (Tony Thorpe). Slang users know how to adapt their language to the context in question.
	14/02/17	Why UK grime artists are staying true to their regional roots.	British grime artists remain loyal to the local accent, and they do not adopt an American one. They make use of a particular accent to construct their identity.
	27/11/19	Birmingham and African caribbean accents face worst bias in UK, study finds.	The article reports the results of a study on prejudices against particular accents conducted at Queen Mary University. MLE receives lower ratings than other accents.[9]
Agence France Presse	26/02/16	Sick, bad, wicked: London's colourful slang on the rise.	J. Green believes that speakers of MLE are not governed by race, class or colour but by age. The variety of English spoken in London could show the way English could evolve in the future. Some artists, who are also speakers of MLE, are proud of the way they speak because they have their own code, and form a family.
Express on line	29/09/16	Queen's English to be wiped out from London 'due to high levels of immigration'.	Immigration is a problem that is affecting the English language.
The Sunday Telegraph	02/10/16	I fink this is the future- but it's just nt proper; in London, the capital of the English-speaking world, the writing is on the all for the sound.	Negative reactions towards MLE which is described as "an egalitarian porridge of mangled consonants, glottal stops, online abbreviations, street slang, gamers' insults, pop lyrics and quotes from the Simpsoms."
BBC Radio 4	14/07/18	Multicultural London English.	It records parts of an interview with R. Drummond on this sociolect. It shows how language is changing. MLE is spoken by young people in the Home Counties.
	12/09/19	Multicultural London English.	James Massiah claims there is no right or wrong way of speaking, but there is a language barrier between different groups of people.
Plus Media Solutions	23/11/18	What two French words can teach us about social change.	MLE reflects the perceived prestige of Jamaican-influenced English among (largely) young people, but it is spoken by people of all ethnicities.

Table 2: (continuation)

[9] This corresponds to the study conducted by Cardoso *et al.* (2019), reviewed above (Section 3).

When considering the views and opinions on MLE as conveyed in these sources, we clearly note that negative attitudes prevail over positive ones. This is something which was expected, and which confirms previous findings (Kerswill 2014; Gates and Ilbury 2019; Cardoso *et al.* 2019; Kircher and Fox 2019a, 2019b; Levon *et al*. 2021; Sharma *et al.* 2022). The positive judgements tend to be seen in contributions from academics and linguists, specifically Rob Drummond, Tony Thorpe and Jonathon Green, whose interest in MLE is mainly linguistic, and highlight the innovative and creative nature of slang. They see MLE as a variety of its own and emphasise the importance of the factor of youth in language innovation and change (BBC Radio 4, 14/07/18; *The Independent*, 05/01/16). Furthermore, MLE is seen as not being conditioned by race (white, black, Asian, etc.), social class (working class versus high class), speaker's area or location (inner London versus outer London) or ethnicity, but only by age, and is considered to be spoken by all ethnicities (*Agence France Presse*, 26/02/16). Also, artists and poets such as Dizraeli and James Massiah consider it as "cool, crazy and rich" and as a group identity marker, with the question of persevering identity appearing here to be crucial (*Metro*, 25/09/15; *BBC Radio 4*, 12/09/2019).

By contrast, the negative assessments of MLE are versed in terms of the same notions reported in previous studies. MLE speakers are regarded as performers and as adopting an artificial accent (*The Daily Star*, 14/03/11). The fact of having so many immigrants in London is seen as negative, with undesirable consequences for the English language, and thus constituting a serious problem (*Express online*, 29/09/2016). Several contributions also claim that MLE is responsible for the displacement of Cockney English, which may disappear within a fifty-year timeframe together with British values more broadly (*The Evening Standard*, 01/02/11). In a similar vein, MLE is considered to be a kind of disease infecting children, with serious consequences for their education and for their future job prospects (*The Daily Mail*, 11/10/13). Even when an academic study conducted by researchers from Queen Mary University on the perception of English accents is reported in the press, emphasis is on the low valuation given to MLE in sharp contrast to RP, French-accented English and Edinburgh-accented English, these being the most highly rated (*The Independent*, 27/11/19). It is difficult to anticipate exactly how information of this kind will be received by the general public and hence how it will influence public opinion, and for this reason the next two sections will explore the perceptions of MLE on social media.

5.3. MLE in social networks: The case of Twitter

The data reported in this section can be regarded as a preliminary survey since it focuses on only one of the social networks, *Twitter*, and thus conclusions should be taken with caution. However, it can help to provide new or additional perspectives on the issue. This preliminary study was conducted in two stages. In the first of these, I considered only the *Twitter* accounts of three rappers (Dizzee Rascal, Wiley, Dappy) for a fifteen-year period (2005–2020). These musicians are generally associated with MLE and use this accent in their speech regularly. The analysis was not restricted to their own posts but also included the responses and retweets of their followers. From the information provided by the accounts of these followers we know that most of them are young adults and are fond of hip-hop, rap and grime music. Some of them are also artists and producers themselves, and the majority of them are based in London. This may explain why lexical and grammar traits of MLE can be easily observed in their exchanges. This is a relevant data for my purposes.

In a second stage, I carried out a similar study but extending the analysis to *Twitter* in general, the only limitation being that the searches and results retrieved all concerned MLE, London English, Jafaican/Jafaikan or Cockney. In this case some of the examples retrieved correspond to extracts from newspaper *Twitter* accounts and other media blogs.

The analysis of the *Twitter* accounts of the three rappers brings together two main ideas. The first has to do with the incorporation of the study of MLE in the English A level curriculum. There are even some tweets that point specifically to the study of the language of Dizzee Rascal, as shown in (18).

(18) We’re studying your language in English atm and are writing an essay about it … wish me luck? (DR 11/12/2014)

No doubt, the incorporation of some features of MLE in the school curriculum of English seems to be a positive policy and may indicate a desire to engender positive attitudes towards this sociolect, in that teenagers will tend to see the academic value of this language as being worthy of study.

The second main idea refers to the influence these rappers are exerting on the English language since they are regarded, by some posters, as precursors of language change and innovation (cf. (19)).

(19) Teenagers in Britain will study Rusty Rockets and Dizzee Rascal as part of a new English A level designed 2 focus on *contemporary use of language* (Ivanka Zonic 08/05/2014)

When considering the tweets attested in the second and wider group of *Twitter* accounts, we also see that views are divided. Some express a preference for the sociolect while others highlight the multicultural nature of this variety and how it has been stigmatised in the media. However, the majority show negative attitudes, believe that the speakers who use this urban dialect sound ridiculous, and that they adopt the accent artificially, as can be seen in (20).

(20) Jafaican may be cool, but it sounds ridiculous. (*Daily Telegraph* blog 29/20/2015)

A set of tweets refer to Cockney and compare it with MLE; most of these allude to the displacement of British values with the emergence of MLE (cf. (21)).

(21) Find it a shame how the cockney accent is slowly disappearing and everyone in London now speak like a fucking roadman (Ben honour 16/06/2017)

Finally, one of the posters calls our attention by mentioning the addition of Jafaican as a new term in the *Oxford English Dictionary* (cf. (22)).

(22) Whateus, chillax, simples, sumfin and Jafaican are some of the new words added to the OED. (*Metro* 16/10/2019)

5.4. YouTube videos on MLE and responses

A total of 11 video documents with their corresponding comments were analysed, amounting to 4,591 comments with an average of 417 comments per video. Overall, the videos can be rated as quite popular since they attained high numbers of views, a total of 2,306,171, with an average of 209,652 per item.[10]

The majority of these documents, which are addressed to the layperson rather than to language specialists, feature the different accents that can be identified under the general category of 'London English', including here classical or traditional RP, Contemporary RP (RP with new developments), Cockney, Estuary English and MLE. In some cases, the presenters illustrate the main differences and, when dealing with MLE,

[10] See Table 3, below, for a full account of the title and website, date of publication, main contents, duration, views and number of comments of the viewers for each of these videos.

they discuss the most relevant pronunciation, lexical and, less often, grammar and discourse features.

These video presentations can be regarded as neutral since the presenters, in general, do not make any critical value judgements about any of these sociolects. They only discuss some of their features. Here is a list of the main MLE features mentioned:

(1) As regards pronunciation, *t*-glottalisation, *l*-vocalisation, ð > *d* thing> ting, θ > f, sharply iambic use of deep voice, etc.

(2) As regards grammar and discourse, high use of address terms (*mate*, *bruv*, *blud*, *man*), use of third person singular present *don't* and negative concord structures, irregular past of BE, invariant tag *innit*, shortening of some words, e.g. *enough*> *nough*.

(3) As regards lexis, the introduction of words having their origin in Jamaican English (*mandem*, *ends*, *yardie*, *yute*, *wagam*, *cotch*) together with other vernacular lexical items (*butters*, *peng*, *safe*, *allow*, *butters*, *bait*, *beef*, *jack*), words undergoing a semantic shift (*sick* meaning cool, awesome), tags with multiple meanings (*innit*, *you get me*).[11]

The comments and reactions included after the videos reveal both positive and negative views towards MLE, although the latter, as before, clearly predominate. In fact, two out of three comments are of a negative kind. Foreigners generally highlight the positive properties of RP and the inarticulateness of MLE, possibly because it does not follow the expected standard, as illustrated in (23).

(23) I'am not british and no native speaker! So maybe I don't get it. But why is this MLE great? Never been to England, but I want to speak this language, as properly as possible, even I don't live there. (Video 1, Learn English with Stormzy. Multicultural London English)

Those who highlight the positive aspects of MLE concentrate on its musicality, its multicultural character, its uniqueness and distinctiveness. Among other opinions, they note it as being a great evolution of Cockney and an effective blend of two cultures, a sexy accent, a positive transformation of the nation's capital embracing multiculturalism,

[11] Notice that some of these features are not exclusive to MLE since they are also present in other London English varieties and even in some general British English dialects. This is the case, for example, with *l*-vocalisation, *sick* as meaning cool, invariant tags as *innit*, negative concord structures, etc.

a posh cockney, that is, cockney with aitches. Its multicultural nature and musicality clearly prevail. (24) to (26) below illustrate some examples.

(24) This video is sick fam! Ha ha. I've been learning MLE from Arsenal Fan TV all these years ha ha ha. Again this video is briliant! Keep it man. (Video 1, Learn English with Stormzy. Multicultural London English)

(25) It's exciting. (Video 2, MLE or Jafaican. BBC1)

(26) I love the multicultural London accent aha. (Video 2, MLE or Jafaican. BBC1)

Negative comments, on the contrary, identify it with the death of English, describing it with the following adjectives and expressions: *non-educated, lazy, ugly, vile, ghastly, horrible, barbaric, chavvy, disgusting, London pidgin, fake, fashionable, failings of multiculturalism, gay version of the original, incorrect/w*rong *way to speak English, trash teen talk, the ebonics of the UK, dumbed down English, lowlife slang, sounds like tramps, horrible accent*, especially hearing it from white people. Multiculturalism is even regarded as a cancer to UK with homophobic and racist overtones here included. Examples (27)–(30) can be regarded as an illustration of this.

(27) The Multicultural accent is the British version of Thug/Gangstas Rap very barbaric. (Video 5, London Accents: RP/Cockney/Multicultural London English)

(28) The MLE is like cancer to my ear. It's associated with low life, aggressive things. It's vile and ghastly. It would me more appropriately called London pidgin, chavvy, disgusting accent, dumbed down English, lowlife slang. (Video 6, London accent tips).

(29) Honestly MLE accent sounds like gay version of the original London accent. (Video 5, London Accents: RP/Cockney/Multicultural London English).

(30) MLE is so far the ugliest accent O have ever Heard. Multicultural means actually White people trying to sound Jamaican. (Video 2, MLE or Jafaican. BBC1).

Overall, the views expressed by the participants focus on the same issues as mentioned above. They highlight its lack of correctness, the negative condition of multiculturalism, its broken and ugly nature, and its association with teen and black speakers. They also regard it as uneducated speech, as common among young speakers, and as not suitable for school and academic purposes. Some views also draw a connection between this accent and lower working class. They also refer to the need to adapt their speech to the

context in question. This means that in their judgements they combine social, educational, racist, and even sexual orientation arguments and criteria.

Title and website	Date	Main contents	Time	Views	Comments
Posh British Girl Teaches Londoner How to Speak English <http://www.youtube.com/watch?v=agV7XYGhFu8>	September 2019	This is an educational video which explores variations with RP and MLE, and how Britain's division of social classes has a bearing on accents. Speakers tend to adapt their language according to the situation, for example, in a job interview.	24:38	14,808	144
Learn English with Stormzy. MLE: <https://www.youtube.com/watch?v=1MQdEVo6Yc>	April 2019	The presenter introduces new features of MLE by illustrating examples from an interview with Stormzy, a British rapper, singer and songwriter.	12.19	57,493	381
London Accents: RP | Cockney | Multicultural London English <https://www.youtube.com/watch?v=_H8r2Izzo5k>	February 2018	They describe what they call London accents. MLE is featured as the newest of the accents heavily influenced by African and Asian communities of speakers and is considered as the most widely used in London.	12:10	555,830	1,139
London Dialects <https/wwww-youtube.com/watch?v=HOQUnt5h8w4>	January 2018	Five different London accents are distinguished. MLE or Jafaican is regarded as the variety of the hip-hop generation, invented by some hipsters and teenagers hanging out.	03:43	29,345	81
Multicultural London English: Dialectable Episode 1 <https://www.youtube.com/watch?v=BKHczYBW6DI>	September 2017	The presenter describes MLE in rather neutral terms. He, first of all, explains how MLE was formed and then refers to distinctive features of this variety.	4:51	7,557	12
London Accent Tips. MLE. *Bruv. Innit. Ting*! <https://www.youtbe.com/watch?v=iUjMmwxmOnY>	June 2017	The speaker presents the video as a tribute to London after the riots that took place in August 2011. It is defined as an amalgamation of the different accents of London that came together.	03:14	62,151	175
How to Talk like a Real Londoner <http://www.youtube.com/watch?v=PbCiNdAAUM4&feature=youtube>	January 2017	MLE is described as having its own rules of pronunciation and grammar. It is a style of English. MLE has replaced Cockney.	13:01	541,451	463
Sick, Bad, Wicked: London's Colourful Slang in the Rise <https://www.youtube.com/watch?v=91Zq0YHxHfg>	February 2016	MLE is regarded as a new variety that is rapidly spreading and with a strong influence from hip-hop. It is also described as an accent that is governed by age and not by race or colour.	01:46	743	0

Table 3: Overview of the *YouTube* videos on MLE considered in the analysis

Title and website	Date	Main contents	Time	Views	Comments
MLE or Jafaican. BBC1 <https://www.youtube.com/watch?v=0KdVoSS_2PM>	May 2015	MLE is described as gaining ground to Cockney. Several features of MLE are described and illustrated with examples.	11:17	55,609	311
The Best British Street Slang <http://www.youtube.com/watch?v=9Z8JqutRWrs>	April 2015	It focuses on MLE. Who speaks it? Ali G, D. Rascal, N-Dubz, hip-hop, grime and garage artists and musicians. Some examples of characteristic words and expressions are provided as examples.	10:26	957,677	1840
Who's an Eastender now? (Paul Kerswill) <https://www.youtube.com/watch?v=hAnFbJ65KYM&feature=emb_title>	September 2011	Part of a general talk delivered by Kerswill who refers to how migration has transformed Cockney. He also alludes to the riots in London and to the evolution of different London varieties. Then he analyses the views of the well-known journalist, Starkey, on MLE who claimed MLE was a foreign variety associated with the black community. In Kerswill's view, Starkey is totally wrong.	18:16	23,507	45

Table 3: (continuation)

6. Engendering positive attitudes towards MLE

As a secondary aim of this paper, I also sought to provide some reflections on measures and initiatives that could be taken to fight some of the stereotypes commented on throughout this paper, towards fostering more open attitudes of respect and tolerance to MLE and its speakers. These reflections could even be extrapolated to other accents which, like MLE, may be stigmatised or regarded as inferior to other varieties. As several scholars have pointed out in a broader context of language change (Trudgill 1975, 1983; Cheshire 1982; Edwards 1984; Cheshire and Trudgill 1989; Cheshire *et al.* 2017; Gates and Ilbury 2019), it would be necessary to discuss the traditional notion of 'standard English' further, especially considering the evolution and diffusion of English nowadays; the same would apply to the notion of 'linguistic diversity'. The introduction of extracts for discussion and consideration from MLE artists and influencers in the A level curriculum seems to be a positive initiative, since it might help towards a fuller recognition of this sociolect and of other varieties which do not necessarily follow what is generally regarded as the standard. In addition, this would be directly connected with one of the learning outcomes of the *Assessment and Qualifications Alliance* (AQA) English syllabus for the A-level in English Language in 2023, which makes reference to the specific "study of social attitudes to, and debates about, language diversity and

change" as well as to the analysis of different texts using different sociolects (occupational groups, ethnicity, gender) and texts using different dialects (regional, national and international). In the learning outcomes referred to in module 3, "Language in Action," specific reference is also made to research projects that could be undertaken regarding how people feel about language.[12] In this respect, we might bear in mind the results of previous studies (Snell and Andrews 2017) that have clearly shown how the inappropriate pedagogical treatment of regional variation can have negative effects on students' educational achievements. Students need to be taught how to switch from their own variety to standardised varieties of English according to the situation in question and this passes necessarily through the contact, appreciation and understanding of these varieties of English and their own mode of expression.

Teachers and educators can also play an important function here by being trained on how to respond to students' own variety and how to deal with all these issues, that is, language attitudes and ideologies, and accent bias in the classroom. However, we should not overlook the role of parents, who can also have an influence on their children. Explaining to them some of the decisions taken in the English classroom and the reasons underlying those decisions could have beneficial effects. The creation of suitable resources and materials with particular attention to MLE and other non-mainstream varieties for their use in the English classroom might also play an important role in this direction. Mass media should also pay more attention to the importance of language diversity and make a positive contribution here, rather than adopting critical attitudes which frequently engender unjustified stereotypes. Students should also be cautioned about the information available on social media regarding attitudes to language and language ideologies, so that they may be in a position to be critical and not to accept everything that is being claimed without reflecting about it, and that, where necessary, they should be able to contrast and question the data.

[12] Further information at: https://www.aqa.org.uk/subjects/english/as-and-a-level/english-language-7701-7702/subject-content-a-level

7. FINAL WORDS

This paper has contributed to the study of attitudes towards MLE and its speakers by providing new data extracted and analysed from corpora (LIC and MLEC), mass media, and also from the social networks *Twitter* and *YouTube*. The latter have turned out to be rich sources of information for the study of language attitudes since they collect large samples of spontaneous thoughts and beliefs, and provide additional perspectives on language attitudes, which may be different from those found in printed material and speech data. It is true that they also show some limitations, especially if compared with corpora-derived data and other methods of attitudes linguistic research (Kircher and Zipp 2022), such as scales, questionnaires, interviews, focus groups, association tests, completion of specific tasks, in that the latter can be regarded as more rigorous and scientific. With data from social media, by contrast, it is not always possible to control closely some of the variables pertaining to the posted comments, with contributors often using nicknames and providing very little information about themselves, thus being difficult to categorise.

In terms of the degree of awareness MLE speakers show regarding their own variety, it was observed that quite often they do not really know how to define it, and that they resort to general labels such as *slang* or *urban speech*. Some younger speakers use the label *Cockney*, although this was not the preferred option by the majority of participants. The age factor seems to play an important role in this respect, since older and white speakers tend to be associated with Cockney, while respondents of the younger generations are more clearly identified with slang or this new urban sociolect.

REFERENCES

Allport, Gordon W. 1954. The historical background of modern social psychology. In Gardner E. Lindzey ed. *Handbook of Social Psychology (Vol 1): Theory and Method*. Cambridge: Addison-Wesley, 3–56.

Augoustinos, Martha, Iain Walker and Ngaire Donaghue. 2006. *Social Cognition: An Integrated Introduction*. London: Sage.

Ajzen, Icek. 1988. *Attitudes, Personality and Behaviour*. Milton Keynes: Open University Press.

Baker, Colin. 1992. *Attitudes and Language*. Clevedon: Multilingual Matters.

Cardoso, Amanda, Erez Levon, Devyani Sharma, Dominic Watt and Yang Ye. 2019. Inter-speaker variation and the evaluation of British English accents in employment contexts. In Sasha Calhoun, Paola Escudero, Marija Tabain and Paul Warren eds. *Proceedings of the 19th International Congress of Phonetic Sciences*. Canberra,

Australia: Australasian Speech Science and Technology Association Inc., 1615–1619.https://accentbiasbritain.org/wpcontent/uploads/2019/10/ICPhSPaper_101218_final_nonAnon.pdf

Cheshire, Jenny. 1982. Dialect features and linguistic conflict in schools. *Educational Review* 34/1: 53–67.

Cheshire, Jenny. 2019. Taking the longer view: Explaining Multicultural London English and Multicultural London French. *Journal of Sociolinguistics* 24/3: 308–327.

Cheshire, Jenny and Sue Fox. 2009. *Was/were* variation: A perspective from London. *Language Variation and Change* 21/1: 1–38.

Cheshire, Jenny and Peter Trudgill. 1989. Dialect and education in the United Kingdom. In Jenny Cheshire, Viv Edwards, Henk Münstermann and Bert Weltens eds. *Dialect and Education: Some European Perspectives*. Clevedon: Multilingual Matters, 94–109.

Cheshire, Jenny, Paul Kerswill, Sue Fox and Eivind Torgersen. 2011. Contact, the feature pool and the speech community: The emergence of Multicultural London English. *Journal of Sociolinguistics* 15/2: 151–196.

Cheshire, Jenny, Jacomine Nortier and David Adger. 2015. Emerging multiethnolects in Europe. *Queen Mary's OPAL (Occasional Papers Advancing Linguistics)* 33 [working paper]. https://www.qmul.ac.uk/sllf/media/sllf-new/department-of-linguistics/33-QMOPAL-Cheshire-Nortier-Adger-.pdf

Cheshire, Jenny, David Hall and David Adger. 2017. Multicultural London English and social educational policies. *Languages, Society and Policy* 1: 1–19.

Clyne, Michael. 2000. Lingua franca and ethnolects in Europe and beyond. *Sociolinguistica* 14: 83–89.

Dragojevic, Marko, Howard Giles, Anne-Carrie Beck and Nicholas T. Tatum. 2017. The fluency principle: Why foreign accent strength negatively biases language attitudes. *Communication Monographs* 84/3: 385–405.

Drummond, Rob. 2018. *Researching Urban Youth Language and Identity*. Basingstoke: Macmillan.

Ebner, Carmen. 2007. *Proper English Usage*: *A Sociolinguistic Investigation of Attitudes towards Usage Problems in British English*. Leiden, The Netherlands: Leiden University dissertation.

Edwards, Viv. 1984. *Language in Multicultural Classrooms*. London: Batsford.

Fox, Sue. 2012. Performed narrative: The pragmatic function of *this is* + speaker and other quotatives in London adolescent speech. In Isabelle Buchstaller and Ingrid van Alphen eds. *Quotatives: Cross-linguistic and Cross-disciplinary Perspectives. Converging Evidence in Language and Communication Research*. Amsterdam: John Benjamins, 231–258.

Fox, Sue. 2015. *The New Cockney: New Ethnicities and Adolescent Speech in the Traditional East End of London*. Basingstoke: Macmillan.

Garrett, Peter. 2010. *Attitudes to Language*. Cambridge: Cambridge University Press.

Garrett, Peter, Nikolas Coupland and Angie Williams. 2003. *Investigating Language Attitudes: Social Meanings of Dialect, Ethnicity and Performance*. Cardiff: University of Wales Press.

Gates, Shivonne M. and Christian Ilbury. 2019. Standard language ideology and the non-standard adolescent speaker. In Clare Wright, Lou Harvey and James Simpson eds. *Voices and Practices in Applied Linguistics: Diversifying a Discipline.* York: White Rose University Press, 109–125.

Herring, Susan C. 2019. The coevolution of computer-mediated communication and computer-mediated discourse analysis. In Patricia Bou-Franch and Pilar Garcés-

Conejos Blitvich eds. *Analyzing Digital Discourse: New Insights and Future Directions*. Cham: Palgrave Macmillan, 25–67.

Kerswill, Paul. 2013. Identity, ethnicity and place: The construction of youth language in London. In Peter Auer, Martin Hilpert, Anja Stukenbrock and Benedikt Szmrecsanyi eds. *Space in Language and Linguistics: Linguae and Litterae*. Berlin: De Gruyter, 128–164.

Kerswill, Paul. 2014. The objectification of 'Jafaican': The discoursal embedding of Multicultural London English in the British media. In Jannis Androutsopoulos ed. *The Media and Sociolinguistics Change.* Berlin: De Gruyter, 428–455.

Kerswill, Paul and Eivind Torgersen. 2021. Tracing the origins of an urban youth vernacular: Founder effects, frequency, and culture in the emergence of Multicultural London English. In Karen V. Beaman, Isabelle Buchstaller, Sue Fox and James A. Walker eds. *Advancing Socio-grammatical Variation and Change. In Honour of Jenny Cheshire*. New York: Routledge, 249–276.

Kerswill, Paul and Heike Wiese eds. 2022. *Urban Contact Dialects and Language Change: Insights from the Global North and South*. New York: Routledge.

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

Kircher, Ruth and Sue Fox. 2019a. Attitudes towards Multicultural London English: Implications for attitude theory and language planning. *Journal of Multilingual and Multicultural Development* 40/10: 847–864.

Kircher, Ruth and Sue Fox. 2019b. Multicultural London English and its speakers: A corpus-informed discourse study of standard language ideology and social stereotypes. *Journal of Multilingual and Multicultural Development* 42/9: 792–810.

Kircher, Ruth and Lena Zipp eds. 2022. *Research Methods in Language Attitudes*. Cambridge: Cambridge University Press.

Levon, Erez, Desyani Sharma, Dominic Watt, Amanda Cardoso and Yang Ye. 2021. Accent bias and perceptions of professional competence in England. *Journal of English Linguistics* 49/4: 355–388.

Lucas, Christopher and David Willis. 2012. *Never again*: The multiple grammaticalization of *never* as a marker of negation in English. *English Language and Linguistics* 16/3: 459–485.

Milroy, James. 2001. Language ideologies and the consequences of standardization. *Journal of Sociolinguistics* 5/4: 530–555.

Mufwene, Salikoko. 2001. *The Ecology of Language Evolution*. Cambridge: Cambridge University Press.

Nortier, Jacomine and Bente A. Svendsen eds. 2015. *Language, Youth and Identity in the 21st Century. Linguistic Practices across Urban Spaces*. Cambridge: Cambridge University Press.

Núñez, Paloma and Ignacio Palacios-Martínez. 2018. Intensifiers in MLE: New trends and developments. *Nordic Journal of English Studies* 17/2: 116–155.

Oppenheim, Bram. 1982. An exercise in attitude measurement. In Glynis M. Breakwell, Hugh Foot and Robin Gilmour eds. *Social Psychology: A Practical Manual*. Basingtoke: Macmillan, 38–56.

Palacios-Martínez, Ignacio. 2015. Variation, development and pragmatic uses of *innit* in the language of British adults and teenagers. *English Language and Linguistics* 19/3: 383–405.

Palacios-Martínez, Ignacio. 2016. Negative intensification in the spoken language of British adults and teenagers. *Nordic Journal of English Studies* 15/4: 45–77.

Palacios-Martínez, Ignacio. 2017. Negative concord in the language of British adults and teenagers. *English World-Wide: A Journal of Varieties of English* 38/2: 153–181.
Palacios-Martínez, Ignacio. 2018. "Help me move to that, *blood*". A corpus-based study of the syntax and pragmatics of vocatives in the language of British teenagers. *Journal of Pragmatics* 130: 33–50.
Palacios-Martínez, Ignacio. 2020. Methods of data collection in English empirical linguistics research: Results of a recent survey. *Language Sciences* 78: 101263. https://doi.org/10.1016/j.langsci.2019.101263
Rampton, Ben. 2015. Contemporary urban vernaculars. In Jacomine Nortier and Bente A. Svendsen eds. *Language, Youth and Identity in the 21st Century*. Cambridge: Cambridge University Press, 24–44.
Rosenberg, Milton J. and Carl I. Hovland. 1960. Cognitive, affective and behavioural components of attitudes. In Carl I. Hovland and Milton J. Rosenberg eds. *Attitude, Organisation and Change*. New Haven: Yale University Press, 112–163.
Sharma, Devyani, Erez Levon and Yang Ye. 2022. 50 years of British accent bias. *English World-Wide: A Journal of Varieties of English* 43/2: 135–166.
Snell, Julia and Richard Andrews. 2017. To what extent does a regional dialect and accent impact on the development of written and spoken skills? *Cambridge Journal of Education* 47/3: 297–313.
Squires, Lauren. 2016. *English in Computer-mediated Communication: Variation, Representation, and Change*. Berlin: De Gruyter.
Torgersen, Eivind, Costas Gabrielatos and Sebastian Hoffmann. 2018. Corpus-based analysis of the pragmatic marker *you get me*. In Eric Friginal ed. *Studies in Corpus-based Sociolinguistics*. New York: Routledge, 176–196.
Trudgill, Peter. 1975. *Accent, Dialect and the School*. London: Arnold.
Trudgill, Peter. 1983. *On Dialect: Social and Geographical Perspectives*. Oxford: Blackwell.
Vessey, Rachelle. 2015. Corpus approaches to language ideology. *Applied Linguistics* 38/3: 277–296.
Wiese, Heike. 2009. Grammatical innovation in multiethnic urban Europe: New linguistic practices among adolescents. *Lingua* 119/5: 782–806.

Corresponding author
Ignacio M. Palacios-Martínez
University of Santiago de Compostela
Department of English and German
Avenida de Castelao, s/n
Santiago de Compostela, 15782
Spain
E-mail: ignacio.palacios@usc.es

received: January 2023
accepted: May 2023

RiCL Research in Corpus Linguistics

Review of Egbert, Jesse, Douglas Biber and Bethany Gray. 2022. *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness*. Cambridge: Cambridge University Press. ISBN: 978-1-107-15138-3. DOI: https://doi.org/10.1017/9781316584880

Javier Pérez-Guerra
Universidade de Vigo / Spain

The aim of this monograph is to provide guidelines and yardsticks, as opposed to definitive rules, to help determine whether the corpus employed for a given linguistic study is representative or not of the type of data being investigated —the reader will note the deliberate use of downtoning expressions in the previous sentence (*guidelines*, *yardsticks*, *help*), reflecting the highly nuanced and uncertain nature of this topic.

The authors, Egbert, Biber and Gray, henceforth EGB, begin by acknowledging the success of corpus linguistics in current linguistic research, which, in Section 1.1, they quantify for us by reporting that corpus-based analyses were used in more than 50 per cent of the 410 papers published in 18 journals in 2014. So, yeah, based on those numbers, you could say that corpus-based/driven 'methodologies' (a term I prefer to 'frameworks', 'approaches' or 'theories') are worth another monograph. This opening section also provides a useful compilation of corpus definitions and concludes that, as linguists, we can agree that a corpus is a possibly large, possibly principled and possibly representative collection of authentic texts, 'representative' being the key word over the next 280 pages of the monograph. (Homework task: select modals and adverbs from the previous sentences to add to the list of downtoning expressions used in the opening paragraph.) If 'representativeness' is the central theme of the study, then the theoretical foundation on which the whole book is built is the principle that corpus linguists analyse linguistic phenomena by inspecting linguistic data in *a* corpus, so every finding or conclusion is circumscribed to *the* corpus we have selected or compiled. Both the 'representativeness'

of the data and the validity of the author's claims are intimately connected: the data are representative of *the* corpus from which the data have been retrieved. That stated and agreed, EBG set themselves the home-by-teatime task of coming up with a formula that will serve to determine that my corpus and, by extension, my findings are representative not simply of *the* corpus but of the language, dialect, period, text type, register, etc. that I am exploring. Conscious that this objective is not exclusive to corpus linguistics, EGB also address the issue of sampling sociolinguistic data and ascertaining representativeness in other population types.

Section 1.3 examines two key factors affecting the multidimensional concept of representativeness: the concepts of 'domain' and 'distribution'. Domain representativeness tells us whether *the* corpus reflects the language, period, register, etc. we want to analyse. Distribution representativeness determines whether *the* corpus is a valid source to scientifically investigate the linguistic phenomena or features of our project. Domain and distribution representativeness must be on the table when we compile (design) and select (evaluate) *the* corpus, and when, as 'corpus consumers' (see Section 1.4), we assess the findings of others based on *a* corpus. I should point out here that the authors employ an initially frustrating but actually brilliant technique of introducing seemingly vital aspects of their proposal in passing (even disruptively) early on and then dropping them for whole chapters, before picking them up again much later in the book. Perfectly illustrative of this are the concepts of 'domain' and 'distribution', which we discover later are central to EBG's notion and calculation of corpus representativeness.

Chapter 2 reviews the different conceptions of representativeness within corpus linguistics. Just as Chapter 1 deals with the different definitions and characterisations of corpus, here EBG document the vast array of ways in which the term representativeness is used. Of the ten uses summarised in Section 2.1, let us focus here on four:

(i) "absence of selective forces", i.e. a "'hands-off' approach to text selection and collection" (p. 31);
(ii) illustrative of "typical or ideal cases" (p. 33), balanced or "proportional of the population's heterogeneity" (p. 34), associated with a 'stratified' corpus design, and "permitting good estimation" of quantitative parameters in the larger population (p. 35);
(iii) "designed for a particular purpose" or function (p. 36); and

(iv) size, based on the premise that "a very large corpus is a *de facto* representative corpus" (p. 36).

The notion of representativeness proposed by the authors in the monograph is thus the sum of the features of these and the other meanings of the term, which are explored in their respective subsections.

Chapter 3 offers an introduction to the "decidedly complex and multifaceted construct" of corpus representativeness (p. 53) as a gradient continuum which should be understood in terms of *greater* or *lesser* representativeness, rather than a "dichotomous, all-or-nothing" notion of perfectly representative versus unrepresentative objects (p. 62). Figure 1 below, which is an adaptation of the authors' Figure 3.1 on p. 54, illustrates and summarises graphically the different factors involved in this continuum.

Linguistic RQs		Domain description Domain boundaries	External information (web, surveys, publications) Internal information (texts, researchers' experience)	
↓		Domain categories	Demographic (gender, education, social status) Situational (mode, setting, topic, purpose)	
	Target domain		Text selection	No selection bias No coverage bias Unambiguous domain categories Feasible sampling
			Sampling	-Proportionality -Randomness
	Distribution	Linguistic variables and levels Size		
Representative corpus				

Figure 1: EBG's framework of representativeness

The first factor that has to be taken into account when determining the representativeness of *the* corpus is the linguistic research goal. This means that representativeness is target-related or, in other words, effective for the analysis of specific linguistic phenomena. The second level of the definition brings in the concepts of 'domain' and 'distribution', mooted in Chapter 1 and explored at length (at last) in Chapters 4 and 5, respectively. As regards the two-fold goal of the monograph (designing and evaluating corpus representativeness), the implementation of the features leading to a representative corpus is explained unidirectionally from the perspective of corpus creation or design. According

to the authors, researchers reflect on domain issues, select the texts and compile a corpus which meets the requirements for representativeness.

Chapter 4 focuses on the subject of domain. Firstly, to describe the domain, to define the domain boundaries, and to establish the domain categories, researchers may use information disseminated in external media (web, publications), surveys carried out with expert informants or simply with language users, their linguistic experience and their own analyses of texts from the domain. British English, novels and text messages are examples of domains, whereas 'domain categories' are defined after the application of demographic (age, gender, education, and socioeconomic status) and/or situational (mode, setting, communicative purpose, and topic) variables. Secondly, the domain must be *operationalised* via a set of texts that can be sampled. The texts have to reflect the range of variation in the domain with no coverage bias, represent the domain categories in an unambiguous way, and be "feasibly sampled to create a corpus" (p. 93). As EBG put it, domain operationalising "should represent not only what is real but also what is realistic" (p. 94). Thirdly, the texts that have been selected are sampled to produce a set of objects that shapes the corpus. The authors introduce the notion of data 'stratification', i.e., data sampled from texts that represent the demographic and situational domain categories, which gives rise to two additional issues: 1) proportionality of the sample with respect to the inventory of domain categories (e.g., same size of sampled texts produced by male and by female speakers), and 2) random sampling, according to which researchers randomly select a number of objects either within the entire operational domain (e.g., random selection of texts written in British English) or from each 'stratum' (or domain category level, e.g., random selection of texts produced by female writers), or simply add to the corpus all the linguistic productions they have been able to collect (e.g., with very specific text categories such as job interviews).

Chapter 5 examines the question of distribution. Here, the optimal design of the corpus is affected by the linguistic variables to be investigated. Whereas the first phase of the design process focuses on selecting corpus objects that provide a reliable image of the domain (e.g., the corpus is valid for research in British English), in this second phase the corpus designer has to consider the distribution of the levels or values of the linguistic variables across the texts in the corpus. The distribution of the variable levels is measured by statistical metrics of accuracy. In this respect, researchers need to be aware of 1) countable items, such as tokens (linguistic forms, e.g., overall number of nouns, words,

syllables), 2) of types (distinct linguistic forms, e.g., different nouns, words, syllables), and 3) of the size of the samples and the corpus. In other words, the corpus has to accommodate enough tokens and types, contain sufficient data to reveal desirable statistical effects, and not be undersampled. Determining the size of the corpus for a given domain, a set of domain categories, and a list of linguistic variables is not an easy task. In Sections 5.4.1 to 5.5, EBG describe well-known statistical measures that help assess the precision of the data and the corpus by quantifying the extent of the variation in repeated applications of the same sampling procedures (pp. 123, 130ff): standard deviation, tolerated confidence intervals of the results, standard error of the sample means, relative standard error of the linguistic variables, saturation, and ceiling effect. The basic idea is that corpus designers (and evaluators) should use statistical tools that help determine whether the size of a corpus is suitable for conducting research in a specific domain, operationalised according to a set of domain categories, based on a number of linguistic variables. The statistical analysis of the corpus data reveals if the corpus is large enough to accommodate a significant number of tokens and types, where *token* and *type* are not restricted to lexical forms but refer to levels or values of the linguistic variables under investigation. To give an example, in my own research on double comparatives (*more cleverer*) in World Englishes (my domain), I not only measure the precision of the frequencies of the monosyllabic and polysyllabic adjectives that are pervasive in English but also that of the occurrence of the tokens representing my variable levels (e.g., *cleverer*, *more clever*, *more cleverer*).

Chapter 6 brings together domain and distribution, and puts the statistical notions and metrics introduced in the preceding sections into practice. In Sections 6.1 and 6.2, which would benefit from neater organisation to avoid a certain circularity in the authors' discussion of the same ideas, EBG add new empirical concepts associated with representativeness, of which the concept of 'parameter estimation' is the most crucial one. Parameter estimation allows us to compare the quantitative distribution or frequency of a variable level in the sample and to determine how well its frequency represents the distribution of the same level in the domain. Precision (discussed in Chapter 5) and parameter estimation may be distorted by faulty designs and lead to biased corpora because the texts in the corpus do not reflect the set of texts required by the operational domain ('selection bias') or because of differences between the domain and the type of texts entering the operational domain ('coverage bias'). The remainder of the chapter

consists of a description of experiments measuring the suitability of corpora for specific linguistic studies. To give a few examples, in Section 6.2, EBG measure mean scores for a number of part-of-speech categories (nouns, adjectives, prepositions, verbs, etc.) in different samples of very large corpora (e.g., the whole of Wikipedia, which constitutes the whole domain) and calculate differences through Cohen's *d* values. The sampling of the large corpus is carried out using a range of techniques: randomised selection, non-random alphabetical selection, equal-size samples within each stratum (e.g., people, sports, films/TV, music). The main conclusion is that selection bias can only be overcome by the application of robust data sampling methods. Contrariwise, the implementation of uncontrolled sampling methods and the design of a corpus with a faulty understanding and operationalising of the domain inevitably lead to findings that are not representative of the pursued domain.

Chapter 7 departs from the more theoretical approach of the preceding chapters and presents the reader with a step-by-step guide to representativeness in both corpus compilation and corpus evaluation. The basic phases or steps are much alike for both: establish the linguistic research questions, specify the domain, evaluate the operational domain, define the linguistic research variables, assess the size of the sample, and carry out experiments to test precision, accuracy and lack of bias. In Section 7.3, the authors illustrate the two processes by designing and evaluating a *Corpus of Yelp Restaurant Reviews* and a *Corpus of YouTube Vlogs*, and outline the statistical tasks required to determine optimal sample size based on the mean distributions of part-of-speech categories and stylometric measures (e.g., word length, type/token ratio), standard deviation and confidence interval ranges. Section 7.4 focuses on evaluating existing corpora, namely the academic subcorpora of the *British National Corpus* (BNC 2007) and the *Corpus of Contemporary American English* (COCA; Davies 2008), as "candidates for a study of academic research writing" (p. 201). The authors describe the operational domain (boundaries: textual sources, period; strata: publication types, disciplines) in each subcorpus, compare both of them through statistics of linguistic variables or parameters that are considered relevant to academic writing (e.g., distribution of premodifying nouns and of noun complement clauses), and report their strengths and weaknesses as far as representativeness of academic writing is concerned.

In terms of the formal features of *Designing and Evaluating Language Corpora*, the chapters of the book also include metadata in the form of boxes with extracts from

publications and comments by the authors. Each of the chapters is prefaced by a one- or two-page summary, which explains the key concepts and ideas to be discussed in the sections that follow. Finally, each chapter in the monograph features exercises and discussion points addressed to the different types of target reader: corpus designers (builders, compilers), corpus analysts (including *butterfly* and/or *armchair* researchers; see Fillmore 1992) and corpus consumers. Although these tasks are not, in my opinion, one of the book's strengths, they are a useful way of reinforcing understanding of the contents and a possible teaching resource for those of us with students to initiate into the mysteries of corpus linguistics. In keeping with the increasing emphasis on corpus design as EBG's methodological account progresses, most of the exercises are aimed at this audience type.

As regards the end section of the monograph, the authors include a useful four-page glossary of the main terms used, references, an index and two appendices, containing, respectively, examples of articles describing stand-alone corpora and a survey of corpora, potentially representative of the English language, which have not yet been evaluated empirically for representativeness. The survey in Appendix B comprises 25 widely-used and relatively large and well-documented corpora[1] which are intended for a wide range of linguistic purposes, and five relatively small and less well documented corpora serving more specialised purposes. The features examined include plausibility of corpus name, date of creation, size (either static or monitor corpora), statement of research goals, domain (general language, varieties, both), full texts versus samples (and sampling techniques: randomness, proportionality), documented operational domain, stratification, sampling, etc.

EBG's monograph is well documented, with all the bibliographical references that readers would expect to find in a serious, up-to-date work of corpus linguistics research: studies on corpus methodologies and linguistic issues based on corpus data, and the actual corpora themselves. The authors' definition and characterisation of what a corpus is and their explication of corpus representativeness are seminal, and the examples used in the experiments are well chosen and illustrate the statistical measures and notions clearly and

[1] The list of corpora includes, among others, the *Corpus of Contemporary American English* (COCA; Davies 2008), the *Corpus of Historical American English* (COHA; Davies 2010), the *Corpus of Global Web-Based English* (GloWbE; Davies 2013), the *Corpus of News on the Web* (NOW; Davies 2016), the *British National Corpus* (BNC 2007), the *Brown* corpus (Hofland *et al.* 1999), the *Santa Barbara Corpus of Spoken American English* (SBCSAE; Du Bois *et al.* 2000), the *International Corpus of Learner English* (ICLE; Granger *et al.* 2020), and the *International Corpus of English* (ICE; Kirk and Nelson 2018).

effectively. Appendix B, which describes and evaluates thirty corpora, is very informative, and corpus practitioners will appreciate the combination of theoretical sections and more practical exercises and experiments based on real data. All in all, the authors have succeeded in constructing a unified framework which corpus builders, linguists, and enthusiasts alike will enjoy and benefit from.

Designing and Evaluating Language Corpora is a pleasurable, useful, reader-friendly addition to the canon. The authors guide the reader through the different phases of corpus compilation and evaluation highlighting the need for a clear definition of the research questions and the domain of which the corpus is intended to be representative. However, regarding the key contribution of the monograph —that is, the premise that corpus representativeness can only be evaluated by taking research niche, linguistic variables or predictors and domain into account— the authors acknowledge a central weakness in their framework, namely, that a corpus cannot be classed as representative in statistical terms precisely because *representative* is not an intransitive adjective but requires a complement argument. In other words, only *representativeness of X* can be evaluated. Representativeness, they conclude, is therefore an 'intrinsically negative' concept and, as a result, "a representative corpus is never possible" (pp. 39 and 56).

References

BNC Consortium. 2007. *The British National Corpus.* http://hdl.handle.net/20.500.12024/2554.

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

Davies, Mark. 2010–. *The Corpus of Historical American English* (COHA): 400 million words, 1810–2009. http://corpus.byu.edu/coha/

Davies, Mark. 2013–. *Corpus of Global Web-Based English* (GloWbE). https://corpus.byu.edu/glowbe/.

Davises, Mark. 2016–. *The Corpus of News on the Web* (NOW). https://www.english-corpora.org/now/

Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson and Nii Martey. 2000. *The Santa Barbara Corpus of Spoken American English*. Philadelphia: Linguistic Data Consortium.

Fillmore, Charles J. 1992. Corpus linguistics or computer-aided armchair linguistics. In Jan Svartvik ed. *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm*. Berlin: Mouton de Gruyter, 35–60.

Hofland, Knut, Anne Lindebjerg and Jørg Thunestvedt. 1999. *ICAME Collection of English Language Corpora*. Bergen: The HIT Centre.

Granger, Sylviane, Maïté Dupont, Fanny Meunier, Hubert Naets and Magali Paquot. 2020. *The International Corpus of Learner English*. Version 3. Louvain-la-Neuve: Presses universitaires de Louvain.

Kirk, John and Gerald Nelson. 2018. The International Corpus of English Project: A progress report. *World Englishes* 37/4: 697–716.

Reviewed by
Javier Pérez-Guerra
University of Vigo
Faculty of Philology and Translation
Department of English, French and German
36310. Vigo
Spain
E-mail: jperez@uvigo.gal

RiCL Research in Corpus Linguistics

Review of Gandón-Chapela, Evelyn. 2020. *On Invisible Language in Modern English: A Corpus-based Approach to Ellipsis*. London: Bloomsbury. ISBN: 978-1-350-06451-5. https://doi.org/10.5040/9781350064546

Arja Nurmi
Tampere University / Finland

This volume aims to bridge earlier, mostly theoretically based research of ellipsis in English with a corpus-based study of Late Modern English. The goals of these two strands of linguistics are somewhat different, as theoretical studies discuss what is possible in a language while corpus-based studies are more focused on what is typical and what patterns of variation can be observed. These differences of purpose make the dialogue of the two approaches challenging at times, but nevertheless valuable. For topics such as ellipsis, where the bulk of earlier work has a more theoretical focus, studies such as the one at hand are of particular merit. This is something Gandón-Chapela does not always seem to see the value of herself, as she is at times almost apologetic for engaging in empirical work (p. 139). The focus on the history of English is equally valuable, as this is a phenomenon still not frequently studied in the historical stages of English, with the exception of Warner (1993, 1997), Nykiel (2006, 2015) and Gergel (2009).

The introductory chapter of the volume contains an extensive discussion of the characteristics of ellipsis mentioned in previous research. The specific focus of the volume is on Post-Auxiliary Ellipsis, divided into two subcategories, Verb Phrase Ellipsis and pseudogapping. The description of earlier points of view starts with the standard reference grammars (Quirk *et al*. 1985; Biber *et al*. 1999; Huddleston and Pullum 2002), and continues with the frameworks of Systemic Functional Grammar, Transformational-Generative Grammar and psycholinguistics.

The second chapter introduces the method, that is, corpus linguistics, as well as the data used. The corpus studied, the *Penn Parsed Corpus of Modern British English* (1700–1914; Krock *et al*. 2016), is somewhat misleadingly named by its compilers and this leads the author of the volume to alternate the terms Modern English and Late Modern English as synonyms, when in fact Modern English encompasses both Early and Late Modern English and starts from 1500. The second chapter includes a careful description of the search algorithm used, and provides useful information for anyone else intending to study ellipsis in the Penn corpora. Gandón-Chapela has also persisted in creating ways to work around the inevitable tagging and parsing errors in the corpus. This chapter is concluded by a careful description of the complex analysis schema.

The third chapter presents the analysis of the results in a wealth of detail, at times overwhelming, but at the same time valuable precisely because of the meticulous description of the variables and the results. In this chapter also previous corpus-based research on ellipsis in Present-day English is integrated in greater detail, as Gandón-Chapela compares her findings to those of other empirical studies. There is great merit in the broad scope of analysis, including grammatical, semantic and discursive variables, usage variables concerning diachrony and genre, as well as processing variables in terms of lexical and syntactic distance. The analysis scheme and the careful study of each variable brings forth new information about possible structures and their frequency. While there is ample data for Verb Phrase Ellipsis, the instances of pseudogapping are rarer. This leads to some discussion of variation that is not statistically significant, but seems to be treated as such anyway (p. 127). The significance testing of results is somewhat sporadic, and it is not always clear why there is testing for some variables and not others.

There are many smaller quibbles one might raise, from counting the archaic second person singular inflected forms (*shouldest*, *shalt*) separately from the other forms of the same verbs as licensors of ellipsis (p. 256–257), or not considering the overall frequency of various modals when looking at their function as licensors (p. 177), given that *will* and *would* are considerably more common than *must*, and the corpus would have provided this point of comparison. Similarly for connectors, it would have been interesting to know how common the investigated connectors are in the data altogether. That is, how far they are specifically connected to ellipsis and how far they are high up the list just because they are frequent (p. 135). At times Gandón-Chapela's focus seems to be more on what

is possible, that is, more theoretical, than what is common in the corpus, as she comments in great detail on the individual examples representing types only rarely attested in the data, but that is obviously a justifiable position also in a corpus-based study. One of the merits of the volume is that it clearly points out gaps in corpus-based research on individual variables of ellipsis.

There is also much merit in the wealth of detail. While it is difficult to see the wood for the trees at times, the discussion of all the aspects of ellipsis and provision of details with numerous examples is obviously highly useful for those wishing to carry out further study with different data sets. It is nice to see the author also bring new linguistic features, such as clause type, to the discussion of ellipsis. Once again, it would have been interesting to relate the frequency of the clause types with ellipsis to the overall frequency of them in the corpus, but as this would have disrupted the focus of the study and considerably added to the workload, it is perfectly understandable this avenue was not pursued.

The discussion of genres and the frequency of ellipsis in them suggests, as the author points out, that the phenomenon is typical of texts related to 'oral', 'spoken' or 'colloquial' language, to use the terms from Culpeper and Kytö (2010: 16). Another way of thinking about this might be to apply Biber's (1988) dimensions, particularly in terms of involved vs. informational texts. It might well be that ellipsis is at home with linguistic features associated with involved texts, and this might help target future corpus-based studies of ellipsis towards texts representing such genres. Biber and Finegan (1997) identify particularly drama but also to some extent letters as consistently representing involved features in historical texts.

The third chapter is concluded by a summing up of the findings in terms of the different variables studied, with a focus on Late Modern English in particular. This is followed by the final chapter, which gives an overall summary of findings and lays out suggestions for further research. One final highly useful feature of the volume follows in Appendix 1, which lines out the corpus tool used, *Corpus Search 2*,[1] as well as the query language and its functions. Appendix 2 provides similarly useful information on the labels used for part-of-speech tagging and parsing in the *Penn Treebank* corpora. For anyone

[1] https://corpussearch.sourceforge.net/

not familiar with the model, these provide a useful introduction and necessary support for understanding what has been retrieved through corpus searches.

While the volume has its problems in terms of, for example, significance testing, the range of linguistic and textual variables analysed provides many potential starting points for the further corpus-based study of ellipsis both in the historical stages of English and in Present-day English. Gandón-Chapela herself suggests further studies using the *Penn-Helsinki* corpora of Old, Middle and Early Modern English, which would seem like a fruitful direction, since the corpus used in this study copies its structure from the *Helsinki Corpus*[2] and shares the parsing and tagging model with the *Penn-Helsinki* corpora. The clearly explained search algorithms the author has developed could be put into use very easily and the results would be comparable in a very direct way. As Gandón-Chapela's results seem to suggest that involved texts are a particularly fruitful ground for ellipsis, a study using the *Parsed Corpus of Early English Correspondence*[3] might provide further interesting data.

While the title of the volume is somewhat misleading, as the volume focuses on a specific subtype of ellipsis and a particular period in the history of English, there is a great deal of value in the first corpus-based diachronic study of post-auxiliary ellipsis in English. This study, even with its flaws, provides a good starting point for future research and gives us much detailed information on ellipsis based on both quantitative and qualitative analysis. The results are valuable for both diachronic and synchronic future studies, and seem to provide new information on what is possible as well as what is typical in case of ellipsis.

References

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

Biber, Douglas and Edward Finegan. 1997. Diachronic relations among speech-based and written registers in English. In Terttu Nevalainen and Leena Kahlas-Tarkka eds. *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*. Helsinki: Société Néophilologique de Helsinki, 253–275.

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

[2] https://varieng.helsinki.fi/CoRD/corpora/HelsinkiCorpus/
[3] https://varieng.helsinki.fi/CoRD/corpora/CEEC/pceec.html

Culpeper, Jonathan and Merja Kytö. 2010. *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge: Cambridge University Press.

Gergel, Remus. 2009. *Modality and Ellipsis: Diachronic and Synchronic Evidence*. Berlin: Mouton de Gruyter.

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

Kroch, Anthony, Beatrice Santorini and Ariel Diertani. 2016. *The Penn Parsed Corpus of Modern British English* (PPCMBE2). Department of Linguistics, University of Pennsylvania. https://www.ling.upenn.edu/ppche/ppche-release-2016/PPCMBE2-RELEASE-1

Nykiel, Joanna. 2006. *Ellipsis in Shakespeare's Syntax*. Silesia, Poland: University of Silesia PhD dissertation.

Nykiel, Joanna. 2015. Constraints on the ellipsis alternation: A view from the history of English. *Language Variation and Change* 27: 1–8.

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

Warner, Anthony R. 1993. *English Auxiliaries: Structure and History*. Cambridge: Cambridge University Press.

Warner, Anthony R. 1997. Extending the paradigm: An interpretation of the historical development of auxiliary sequences in English. *English Studies* 78/2: 162–189.

Reviewed by
Arja Nurmi
Faculty of Information Technology
and Communication Sciences
P.O. Box 300
FI-33014 Tampere University
Finland
e-mail: arja.nurmi@tuni.fi

RiCL Research in Corpus Linguistics

Review of Smitterberg, Erik. 2021. *Syntactic Change in Late Modern English: Studies on Colloquialization and Densification*. Cambridge: Cambridge University Press. ISBN: 978-1-108-56498-4. DOI: https://doi.org/10.1017/9781108564984

Bettelou Los
University of Edinburgh / United Kingdom

The Late Modern English period (LModE, c.1700–c.1900) has long been claimed to offer little scope to the study of historical syntax other than shifts in the frequencies of use of syntactic constructions.

> Since relatively few categorical losses or innovations have occurred in the last two centuries, syntactic change has more often been statistical in nature, with a given construction occurring throughout the period and either becoming more or less common generally or in particular registers. The overall, rather elusive effect can seem more a matter of stylistic than syntactic change (Denison 1998: 93).

Smitterberg claims that this alleged lack of innovation and change is at odds with what we know of societal changes taking place in this period: its technological and sociocultural transformations must have produced many more weak network ties (in the sense of Milroy and Milroy 1985: 2–4) than earlier societies, and, if weak ties are assumed to facilitate language change, the picture of linguistic stability claimed for LModE in the literature cannot be correct. The solution to this 'stability paradox' is to move away from a conception of English as a unified whole as the object of the investigation and instead focus on the idiolect as the locus of language change. The texts that make up the historical corpora we rely on for our data are well-known for not being representative of the full range of English speakers, skewed as they are towards the "male, literate, and/or high-status speakers" (p. 4), but it is nevertheless possible to investigate idiolects by proxy if we study the evolution of particular genres.

After an introductory chapter discussing the aims and scope of the study, the second chapter, "Sociocultural and linguistic change in Late Modern English," tackles the two elements of the stability paradox: the increase in weak network ties (resulting from the shift from a mainly rural to a mainly urban society, the fact that social mobility was on the rise, and new modes of travel and communication) and our knowledge to date of linguistic change taking place in this period in lexis, pronunciation, but particularly syntax. This chapter also lays some of the groundwork for changes within genres, in particular earlier work on the development of a distinctive style for academic English with increasing phrasal complexity ('densification') which will be the topic of Chapters 7 and 8. Within this register, there is further diversification in the course of LModE, with the development of distinct stylistic conventions for individual disciplines, like medicine and history. The chapter also reports on earlier investigations into oral versus literate styles, and the appearance of speech-based features in the latter (a phenomenon known as 'colloquialization') on which Chapters 5 and 6 will build further. The third chapter, "Aspects of language change," homes in on the notion of idiolects, and idiolectal change, and the limits of what we can retrieve about past idiolects. Much of the chapter is taken up by a survey of the various positions taken in the literature about the locus of change, and what counts as change —innovation and propagation, or only propagation, that is, we assume that a change has taken place only if it starts to spread. Chapter 4 offers a detailed description of the methodology behind the use of historical text corpora; this chapter does not necessarily offer a new perspective to the readers, rather it validates much what users of such corpora know intuitively, as it makes explicit the justifications for using these datasets. There is often a trade-off between the two important requirements that make historical text corpora suitable tools for studying linguistic change: the requirement that they are representative (i.e. an accurate reflection of the language produced by speakers at the time) and the requirement that the sample for each subperiod is comparable (such that any differences between the output of different historical stages reflect change between these periods rather than epiphenomena due to skewed sampling). There is no one way to operationalize representativeness so that the corpus is similar to the total output of a communal variety. Biber (1993) argues that this could best be achieved by a corpus that mainly consists of conversation, while Leech (2007: 80) claims that texts which have the largest reach in terms of readership should have a prominent place. The problem is, of course, as noted by Váradi (2001: 80), that we do not have enough knowledge of our target population to achieve "fully representative sampling" —if we

did, we would not need a representative sample in the first place! Another topic that is explicitly addressed in this chapter is the issue of what, in scenarios of linguistic competition, counts as variants; this concept, originally used for sociolinguistic investigations into phonological change, is much trickier to implement at other levels of linguistic description, and all the more so in the case of historical work, where investigators cannot rely on introspection to arrive at an adequate inventory of potential variants. A text-linguistic approach with normalized frequencies, treating texts or subcorpora rather than individual tokens as an observation may be a safer choice. The chapter ends with a description of the corpora used for this study: the *Corpus of Nineteenth-Century English* (CONCE; Kytö *et al*. 2006) and the *Corpus of Nineteenth-Century Newspaper English* (CNNE),[1] totaling around 1.3 million words. CONCE contains samples from seven genres (parliamentary) debates, drama, fiction, history, (private) letters, science, and trials; for material drawn from CNNE, Smitterberg focused on two time-spans: 1830–1850 and 1875–1895, on the rationale that it is from 1830 that newspapers start to target a wider range of readers than their traditional educated, high-status readership, while the second period falls within the 'golden age' of newspapers, after the introduction of the telegraph and the telephone revolutionized news reporting. The next four chapters present the study's findings with respect to two ongoing changes: colloquialization (*not*-contraction in Chapter 5 and co-ordination by *and* in Chapter 6) and densification (nouns as premodifiers in NPs in Chapter 7 and participle clauses as postmodifiers in NPs in Chapter 8). All these investigations are models of their kind: extensive data collections, carefully analyzed and documented in terms of what factors are considered and why, and what to take away from the results.

Chapters 5 and 6 argue that there is a trend towards colloquialization on the basis of increasing rates of *not*-contraction and of phrasal versus clausal coordination (here called super-phrasal coordination, as many of the non-phrasal conjoins do not represent complete clauses). The coordination findings are more complex in that there are clear genre differences, and within the letters genre, gender differences: in men's letters, super-phrasal coordination increases, in line with other colloquialization markers, but in women's letters, this particular feature shows a decrease. Smitterberg points to a solution suggested in Culpeper and Kytö (2010: 174) —women's letters retain an older method of text-structuring that does not follow printing conventions for sentence division and

[1] https://varieng.helsinki.fi/CoRD/corpora/CNNE/

punctuation, and uses dashes and super-phrasal *and* instead. Their decrease in super-phrasal *and* over time is due to an increased sense that the sentence rather than the clause should be the basic syntactic unit.

Chapter 7 argues that there is a trend on the basis of increased frequencies of nouns premodifying other nouns, so that we get *telegraph wires* as an alternative to descriptions like *wires that transmit telegraph messages*; *infant son* as an alternative to *our son, who was an infant*; *the police version* as an alternative to *the police's version*; and *ocean life* as an alternative to *oceanic life* or *life in the ocean* (pp. 187–188, 192–193). The varied nature of the longer descriptions —adjectival premodifiers, genitive determiners, and phrasal or clausal postmodifiers— makes tracking the frequencies of the variants virtually impossible, so that a text-linguistic analysis was conducted instead. The results offer a different perspective from previous investigations in that this type of densification is not confined to news and science writing but evident in other genres as well, like drama, fiction and letters, where space is not at a premium. The change, then, is not confined to those genres where it offers a practical advantage, but also incorporates a general shift towards more nouns in the premodifier slot, affecting speech-related writing, and proceeding along the kind of trajectory we might expect for change from below, with women leading the change. This chapter also contains an investigation into what semantic relations can be distinguished between the premodifying noun and the noun head (19 in all; pp. 206–207) and how their frequencies shift over time; unsurprisingly, different genres favor specific types of semantic relations (p. 214). A second investigation focuses on semantic relations with proper names as premodifying nouns.

Chapter 8 focuses on the relative frequencies of participle clauses that function as postmodifiers of nouns. Present-participle clauses often appear to be condensed versions of active relative clauses (*the air passing the windways* vs. *the air that was passing the windways*) while past-participle clauses appear to be condensed versions of passive relative clauses (*a vessel specially built for the purpose* vs. *a vessel that was/has been specially built for the purpose*) (p. 222), so that the frequencies of all four constructions are examined to see whether here, too, we see a trend towards densification. As it is difficult to exclude participle or relative constructions that are not interchangeable from the data, the chapter analyzes participle clauses both from both a variationist and a text-linguistic perspective. The picture that emerges is much more nuanced and much less straightforward than one might expect, and certainly not easily framed in terms of

competing variants. There is support for the hypothesis in that some genres exhibit densification, with restrictive past-participle clauses becoming more frequent than passive restrictive relative clauses in letters, but only because the latter appear to be increasingly avoided. The frequency of restrictive past-participle clauses increases in news (CNNE), by 26 per cent, while that of restrictive present-participle clauses increases by 72 per cent in science —two genres already shown to favor densification in previous research, so that these increases can be argued to be part of the same process, even if they are not matched by a corresponding decrease in the relatives. History and debates, meanwhile, exhibit the opposite trend, of a decrease in past-participle clauses —for debates, this may be due to a shift from indirect to direct speech; for history, it could be the result of its more narrative focus (p. 246), the fact that publications in this discipline tended to be of book-length, and that its readership was less specialized (p. 244). In fiction, non-restrictive present-participle clauses show an increase of 18 per cent. Any scenarios in terms of competing variants are thwarted by the fact that many non-restrictive present-participle clauses are ambiguous between an adnominal and an adverbial reading, and are only equivalent to relative clauses in the former; and also by the phenomenon that prepositional and adjectival phrases postmodifying nouns allow expansion to relative clauses in many cases, too, so that there is, in theory, a larger variant field than just the four constructions examined in this chapter. This probably means that the text-linguistic approach is the safest option here.

The data chapters, then, confirm earlier verdicts of syntactic change in the modern period, like Denison's quoted at the beginning of this review, as a matter of changing frequencies at the level of style rather than true syntactic innovation. The concluding discussion of Chapter 9 gropes towards a solution to the 'stability paradox', which is necessarily speculative. There is stability at the level of the communal language but not at the level of the idiolect, although the changes that emerge at the latter level are subtle rather than drastic. If we see the dissemination of innovations by means of social networks as a social phenomenon of linguistic accommodation, we would not expect truly novel structures to be propagated through that route; novel structures as the product of a single individual that have not diffused to other speakers yet would not be propagated by linguistic accommodation as there are no speakers to accommodate to. The author acknowledges that this line of reasoning —that there is no correlation between social networks and this type of change— leaves unexplained why the rate of truly innovative

change in LModE has slowed down compared to earlier periods. The emergence and dissemination of a standard, and the much higher literacy levels, forcing speakers to acquire an increasingly wide range of usage, may have acted as brakes on innovation.

To conclude, this monograph makes an excellent contribution to the field, is extremely well-written, and a model of research. There are many methodological 'caveats' to present any findings in the right perspective: Figure 2.1, charting lexical innovation, suggests a steep rise in the nineteenth century, but Smitterberg reminds us that the coverage of the *Oxford English Dictionary* is known to be poor for the eighteenth and extensive for the nineteenth centuries (p. 25). At all times, there is an awareness of the type of texts which were used to compile his two corpora, and how they might impact the results (e.g. the discussion of the impossibility of separating out opinion pieces from more neutral news articles in CCNE on pp. 122–123, the discussion of editorial interference influencing rates of *not*-contraction on p. 130, or the impact of specific topics —*meat juice*, *coal cart*— on the frequencies of premodifying nouns in trials, on pp. 199–201). The level of detail provided about the specific social, legal and technological conditions that fostered the growth of newspapers and their readerships adds a great deal of interesting information about how various factors conspired to lead to the emergence of newspaper English as a distinct genre, such as the higher levels of literacy in the general population, the relaxation of libel laws, and the abolition of stamp duty which not only lowered newspaper prices but also facilitated the introduction of the rotary press with its continuous rolls of papers, now that there was no longer a requirement for every single sheet to be stamped (pp. 115–118). Newspaper profits became increasingly dependent on advertisement revenue, requiring market research into readerships, which in turn led to a diversification into different styles for different newspapers, responding to the level of education of the readership they were aiming at.

There were only two occasions where I felt more detail could have been provided. The fact that significance testing was given a section of its own in Chapter 4, and the fact that that section offered a critique of traditional significance testing for historical data (as the null hypothesis assumes randomness by default, whereas language is never random), led me to expect more of a discussion of the alternatives to traditional significance testing, such as Bayesian methods, but this was not forthcoming. When logistic regression models are used in the data chapters, the book is surprisingly coy about details —the term 'glm' is launched on p. 148 without any explanation of what it stands for (generalized linear

model), or that it is part of an *R* package. The second occasion was the discussion of densification by means of premodifying nouns in Chapter 7; I missed references to Halliday's work about the language of science, where a text first introduces and discusses a (scientific) phenomenon, and then uses increasing compression to refer back to the phenomenon once it has been established, so that the compound the writer ultimately ends up with is a one-off formation specifically constructed for a very local purpose, that is, as a referring expression (Halliday 2001: 185; see also Halliday 2004); the same phenomenon has been noted as a morphological innovation by Kastovsky (2006: 207). A hint of this important function occurs on p. 211, where the example *the land question* seems to me exactly this type of one-off. The author discusses the greater effort required by the readership in terms of being able to identify such combinations, but does not make the connection to it serving as a referring expression in the discourse. In contrast, Chapter 8 has much more of an eye for the discourse functions of, for example, non-restrictive past-participle clauses in narratives, in terms of marking backgrounding in fiction.

References

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

Culpeper, Jonathan and Merja Kytö. 2010. *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge: Cambridge University Press.

Denison, David. 1998. Syntax. In Suzanne Romaine ed. *The Cambridge History of the English Language, Vol. 4: 1776–1997*. Cambridge: Cambridge University Press, 92–329.

Halliday, M.A.K. 2001. Literacy and linguistics: Relationships between spoken and written language. In Anne Burns and Caroline Coffin eds. *Analyzing English in a Global Context: A Reader*. New York: Routledge, 181–193.

Halliday, M. A. K. 2004. *The Language of Science*. In Jonathan J. Webster ed. *Collected Works of M.A.K. Halliday: v. 5.* London: Continuum.

Kastovsky. 2006. Vocabulary. In David Denison and Richard M. Hogg eds. *A History of the English Language*. Cambridge: Cambridge University Press, 199–270.

Kytö, Merja, Mats Rydén and Erik Smitterberg. 2006. Introduction: Exploring nineteenth-century English – Past and present perspectives. In Merja Kytö, Mats Rydén and Erik Smitterberg eds. *Nineteenth-Century English: Stability and Change*. Cambridge: Cambridge University Press, 1–16.

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

Milroy, James and Lesley Milroy. 1985. Linguistic change, social network and speaker innovation. *Journal of Linguistics* 21/2: 339–384.

Váradi, Tamás. 2001. The linguistic relevance of corpus linguistics. In Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie and Shereen Khoja eds. *Proceedings of the Corpus Linguistics 2001 Conference, Lancaster University (UK), 29 March–2 April 2001*. Special issue of *UCREL Technical Papers* 13: 587–93.

Reviewed by

Bettelou Los
School of Philosophy, Psychology and Language Sciences
3 Charles Street, EH9 9AD
Edinburgh
United Kingdom
e-mail: b.los@ed.ac.uk

RiCL Research in Corpus Linguistics

Review of Tamaredo, Iván. 2020. *Complexity, Efficiency, and Language Contact. Pronoun Omission in World Englishes.* Bern: Peter Lang. ISBN: 978-3-034-33902-5. https://doi.org/10.3726/b16943

Edgar W. Schneider
University of Regensburg / Germany

This is a really interesting monograph with a precisely defined goal: exploring the intersection of the two topics mentioned in the title, linguistic complexity as evidenced in patterns of pronoun omission in World Englishes. This research design translates a fundamental question in linguistics into an appealing, multi-faceted project: Does language contact result in simplified linguistic varieties? World Englishes are seen as typical manifestations of contact-induced varieties, and the deletion of pronouns in subject or object position, not typically regarded as a characteristic feature of World Englishes, is considered as a model case which allows the investigation of processes of linguistic simplification and restructuring. Both aspects are systematically and comprehensively introduced, and a variety of perspectives and data types selected allows insights into the interrelationship between the two and, more widely, the fundamental questions raised.

Perhaps surprisingly, the "Introduction" starts with an outline of a biological experiment which shows that humans exploit energy maximally efficiently (in walking), and from there it extrapolates in a most appealing fashion to a consideration of language processing, arguing that languages also strive towards an optimal encoding in communication, as shown in 'Zipf's law' and work by Hawkins (2004). The author argues that pragmatic inference contributes to that efficiency, allowing under-specified information to be drawn from context and thus contributing to complexity reduction. Pronouns are introduced as model instances of units whose referents can be retrieved

and which can thus be omitted fairly readily. In the light of contact influences as manifested in different branches of research on World Englishes, pronoun omission is thus argued to be a suitable test case for the interrelationship between complexity reduction and contact. Two convincing research questions are consequently formulated, asking for the contexts and conditions of pronoun omission in the light of efficient information encoding and for the constraints governing this process.

The next two chapters circumscribe the two main topics underlying the study and point out what is known (or postulated) about them from earlier research, and they do so in a highly informative, concise and well readable fashion.

The notion of linguistic complexity and the question of whether all languages are equally complex (with a postulated internal trade-off between structural domains with high complexity and lower complexity in others compensating for this) has been discussed and researched intensely over the last few decades, with a number of classic publications and collective volumes on the issue. The author provides a masterful survey of the history of the notion (with a special eye on the idea of complexity invariance, now largely refuted), different approaches towards and assumptions on it, earlier investigations, and metrics suggested to measure it. Sources of complexity variance, such as the passage of time, language contact, or the role of adult language acquisition, are considered, leading to a discussion of underlying processing principles posited by Williams (1987) and Filipović and Hawkins (2013). Comparing sociolinguistic and typological claims on the issue, it is argued that short-term adult contact reduces complexity but long-term childhood bilingualism increases it. Three proposals on how to measure complexity and associated principles (e.g., economy, transparency and isomorphy, the number of distinctions encoded, or postulated properties of the 'human processor') are discussed and compared in great detail and with sensitivity to the difficulties involved, resulting in a distinction between different kinds (and concepts) of complexity, namely, systemic vs. structural complexity, and also global-local, absolute-relative and overt-hidden complexity. For anybody interested in the complexity debate this chapter offers a recommendable summary. It concludes with a survey of work on the interrelationship between complexity levels and varieties of English, arguing in general that high-contact varieties (language shift varieties as well as pidgins and creoles) are simpler in some respects than low-contact varieties (traditional dialects).

Chapter 3 focuses on pronoun omission, highlighting different approaches to the phenomenon (the generative notion of a pro-drop parameter or cognitive explanations along context and accessibility), earlier findings concerning its occurrence cross-linguistically and in the history and varieties of English and, perhaps most importantly, constraints which have been found to govern the process. Several factors that license pronoun omission are identified, discussed and illustrated, including the presence of agreement morphology, the retrievability of an antecedent, priming effects, verb semantics, coordination, style or chunking effects. Considering some of the complexity distinctions assessed earlier, the author then concludes that pronoun omission represents a case of simplification, since "formal complexity is minimized" (p. 92) with one form less to process, while for the hearer online processing is not made more difficult.

The next two chapters present empirical studies which, although sharing the topic, are completely independent. Chapter 4 scrutinizes data from the *Electronic World Atlas of Varieties of English* (eWAVE; Kortmann *et al.* 2020), which documents the presence or intensity of 235 linguistic features in 77 varieties. Five of the features represent different types of pronoun omission, and these are analyzed with respect to their region —with varieties assigned to eight large-scale world regions which are viewed as a proxy for possible substrate influences, a somewhat problematic assumption given the high degree of multilingualism in many world regions— and variety type (as categorized within eWAVE). These two factors are treated as predictors in a linear regression analysis, with indexes of attestation and pervasiveness as dependent variables, and both sum values and the individual features are considered. The results show that contact varieties (shift L1s, indigenized L2s, pidgins, and creoles) have higher attestation rates than low-contact varieties (traditional and dialect-contact L1s), and region has a significant but weaker impact (merely residual after accounting for variety type) on the pervasiveness of the feature in Asia and the Pacific region. The impact of region is stronger, however, for individual pronoun omission features. A global assessment of these findings, including a comparison to distributions in the *World Atlas of Language Structures* (WALS; Dryer and Haspelmath 2013) confirms these tendencies and supplements a few more details.

Chapter 5 adds a full-blown corpus analysis of the constraints effective in pronoun omission. Three components from the *International Corpus of English* (ICE; Kirk and Nelson 2018) were chosen, namely Great Britain as a low-contact variety, India as a

high-contact L2 variety, and Singapore as a language form in between (also considered high-contact, but with many speakers using English as an L1). This is a decision which is defensible, though it might have been even more interesting to pick an African variety instead of Singapore, which is exceptional in many ways, and perhaps investigate one or two more varieties overall. While pronouns can be retrieved automatically, zero pronouns cannot (which requires a lot of reading and manual searching), so it makes sense to restrict the source texts to a small fraction of ICE. The selection criteria described (p. 136) are convincing. A really interesting methodological decision which might have been argued for a bit more extensively is the selection of an equal number of zero forms as the random sample size of attested pronouns, since in reality pronoun omission occurs substantially less frequently. Thus, while internally the relationship between constraints with and without omission is of interest (and certainly comes out more clearly forced like this), a certain distortion effect cannot really be ruled out (pp. 171–172), since the set and number of omitted pronouns considered, as opposed to attested ones, is boosted. The analysis itself considers two language-external variables and eleven language-internal variables, which are well explained and illustrated, together with hypotheses on their expected impact on pronoun omission. Data presentation then shows both univariate results (i.e., frequencies of present vs. omitted pronouns per variable, presented graphically) and, for complex interaction effects, binary mixed-effects regression modelling as well as random forests per variety, to be able to compare effect strengths.

The results presented in Chapter 5 are rich and manifold, and clearly too voluminous to be reported in detail here. The univariate description of findings is concluded by a nice summary, which identifies a disproportionately high omission rate in Singapore (in my view clearly a substrate effect), and higher omission rates in writing than in speech (which may have to do with more processing time available and easier access to antecedents in written texts). Omission occurs strongly but clearly not exclusively in contexts which reference grammars have described as ‘canonical’ for this phenomenon (like in coordination, clause-initial position, or declarative main clauses). It is also shown that cognitive and processing effects indicate a trend towards efficiency, that is, when omission does not increase the addressee’s processing load in decoding.

The multivariate analysis shows eight predictors and two random effects (speaker and verb form) to have a significant impact. The state-of-the-art statistical machinery

employed is explained by the author in an accessible manner, and goodness-of-fit tests show the model to have strong explanatory power. Boxplots visualize the model predictions and help to make the results accessible. Factors which promote the omission of pronouns include high accessibility of the antecedent, priming (by another instance of omission in the preceding context), main clauses, pronoun reference to other than the speaker or hearer, and lexical and modal (as opposed to non-modal auxiliary) verbs. In addition, a few interaction effects involving variety are worked out; for example, Singaporean English and less so Indian English but not British English tend to omit pronouns also in non-initial positions. Random effects (i.e., different verbs and speakers) also have varying preferences. Overall, as is shown convincingly in the "Discussion" (5.4.1.2), the contact varieties tend to display higher omission rates in more contexts than British English. The findings are compared to earlier claims on the issue and to the author's preliminary hypotheses. Structurally, pronoun omission is evaluated as contributing to simplicity and efficiency, as it is favored in contexts associated with "independent cognitive and processing motivations" (193). Relative system complexities are then measured by running random forests per variety, again yielding strongly predictive model fits. Regional per-variety grammars show some similarities in the ranking of effective constraints (for example, coordination is always the strongest) but also differences in the ranking and number of factors (with Singapore showing nine rather than eight predictors to be effective). The line of argumentation is interesting, though I am not wholly convinced whether the small number of differences is sufficient to attribute to Indian English as the only pure L2 variety the simplest pronoun omission grammar (197), with the same number of constraints effective as in GB.

Overall, the author finds the interaction between contact and complexity to be "intricate" (p. 199), summarizing and highlighting some varying but also some shared tendencies. Pronoun omission is argued to represent structural simplification and to contribute to communicative efficiency, with reduced production and processing efforts. A distinction between structural and systemic complexity is strictly upheld; British English is regarded as the variety with the most complex grammar structurally, and Indian English, as a clear product of L2 acquisition, is claimed to be the simplest systemically. Substrate impact is largely rejected because Singapore's constraint ranking is similar to (and statistically positively correlated with) that of Great Britain.

This is a claim of which I am not totally convinced, since Singapore's proportion and frequencies of omission tend to be highest in many contexts, which, to my mind, is perfectly in line with positing a Sinitic substrate (pp. 208–209). Perhaps the small difference in the number of significant constraints in the random forest analyses (in addition to variable results as to their ranking) is given a bit too much weight as opposed to other findings.

Chapter 6, "Concluding remarks and suggestions for further research," summarizes the author's views and findings on the interrelationship between linguistic complexity and communicative efficiency, as showcased in his thorough investigation of pronoun omission as a model application case. While coming down to a clear preferred baseline interpretation (pronoun omission contributing to simplicity), the author always provides a balanced argumentation. He juxtaposes his finding of "support for the claim that pronoun omission results in simplification without a loss in communicative efficiency, at least in structural terms" (p. 205) with the opposite position that "in systemic terms, pronoun omission produces more complex grammars as it entails a larger set of referential expressions and, possibly, more rules to account for their use" (p. 205). And I agree that one of the impressive and interesting findings of the study is having "uncovered a potential trade-off between structural and system complexity in S[ingapore] E[nglish]" (pp. 206–207).

The author is to be congratulated on having provided a fine, most sophisticated case study which tackles the issue investigated systematically, comprehensively, and with great theoretical and methodological awareness and rigor. His research has creatively combined some theories and branches of linguistics which, as is clearly shown, could profit from more regular and substantial interaction. This is a tightly circumscribed model case study digging deep and offering valuable insights and also a lot of food for thought. Still, understanding linguistic complexity (and even more so in its interaction with contact) simply remains a most complex task.

References

Dryer, Matthew S. and Martin Haspelmath. 2013. *World Atlas of Language Structure*. https://walsh.info.

Filipović, Luna and John A. Hawkins. 2013. Multiple factors in second language acquisition: The CASP model. *Linguistics* 51: 145–176.

Hawkins, John A. 2004. *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Kirk, John and Gerald Nelson. The International Corpus of English project: A progress report. *World Englishes* 37/4: 697–716.

Kortmann, Bernd, Kerstin Lunkenheimer and Katharina Ehret. 2020. *The Electronic World Atlas of Varieties of English*. http://ewave-atlas.org.

Williams, Jessica. 1987. Non-native varieties of English: A special case of language acquisition. *English World-Wide* 8: 161–199.

Reviewed by
Edgar W. Schneider
University of Regensburg
Department of English and American Studies
DE-93040 Regensburg
Germany
e-mail: edgar.schneider@sprachlit.uni-regensburg.de

RiCL Research in Corpus Linguistics

Review of Bouzada-Jabois, Carla. 2021. *Nonfinite Supplements in the Recent History of English*. Bern: Peter Lang. ISBN: 978-3-034-34226-1. DOI: https://doi.org/10.3726/b19142

Patrick Duffley
Laval University / Canada

This monograph explores subjectless *-ing* and *-ed* supplement constructions in the recent history of English from a corpus-based perspective. Supplements are defined as constructions in the clausal periphery that do not fulfil a core syntactic function within the matrix clause, and whose deletion typically does not have syntactic, semantic or grammatical consequences for either the structure or the interpretation of the clause. Despite their peripheral status, supplements are prototypically linked to the main clause in various ways. The analysis of these two very common types of non-finite supplement allows for a better characterization of the periphery of the clause in terms of more and less prototypical elements. The monograph also contributes to the description of the diachronic variation of the features that characterize the construction in Late Modern English and Present-Day English. On this level, the study reveals increasing homogeneity among supplements over time and proposes that this reflects a trend towards the regularization of the non-finite periphery in English.

Chapter 1 introduces the construction which is the focus of the study and Chapter 2 presents the review of the relevant literature and a survey of the main features that characterize it, also providing a terminological overview of the concept of supplement and examining a number of features that have been used to define this concept with a view to establishing a clear-cut definition of the term and distinguishing it from other similar constructions. Chapter 3 deals with methodological issues concerning corpus linguistics in general, the corpora used for the analysis of supplements in the study, as well as the retrieval process used to build the database. Chapters 4 and 5 represent the

core of the study and provide an in-depth analysis of *-ing* and *-ed* supplements in Late Modern English and Present-Day English. The final chapter summarizes the results of the study and proposes possible avenues for future research.

The author is aware that the subject of her study is not easy to define. She proceeds through a rigorous examination of the various tests proposed in the literature for identifying supplement clauses and concludes, quite rightly, that neither the impossibility of clefting (p. 72), nor the impossibility of being the focus of a question (p. 75), nor the fact of being outside the scope of negation (p. 77), nor that of being excluded from verb phrase anaphor (p. 79), are sufficiently reliable diagnostics for identifying such clauses, as shown in (1)–(4) respectively:

(1) a. *Going down a hill*, the horse threw him over his head.
b. It was going down a hill that the horse threw him over his head.

(2) a. *Told of some business that drew her to where he was hiding*, she said she would be glad to help.
b. When did she say that she would be glad to help?

(3) a. *Just staying in the shade*, one does not remain hydrated.
b. One does not remain hydrated just staying in the shade but drinking lots of water.

(4) a. *Used with due care*, this ointment may be applied again and again to the same region of the body.
b. And *so* may this other ointment [= this other ointment may be applied again and again to the same region of the body if used with due care].

She concludes that "the syntactic dependency or integration of supplements with respect to their main clauses is viewed as a scalar property of the construction, in that it involves a continuum from more to less syntactically dependent or integrated supplements" (p. 80). To be included in the database of the study, supplements do have to meet certain criteria however: "they have to show a clearly adverbial reading, be able to move to a position other than post-subject, and be understood as influencing the whole event in the main clause and not just the subject" (p. 86).

Even these minimal criteria are not unambiguously applicable, however. First of all, it is hard to define what a 'clearly adverbial reading' is: for example, if

correspondence to a *how*-question is taken to characterize the prototype of such a reading, the adjective *sick* would have to be analyzed as manifesting an adverbial 'manner' reading in (5) below:

(5) She was sick.

Bouzada-Jabois herself is aware moreover of the difficulty in distinguishing non-restrictive reduced adjectival clauses from adverbial supplement clauses (pp. 44–47). Non-restrictive adjective clauses can be argued in certain cases to meet the third criterion, that of influencing the whole event, as illustrated by (6) below:

(6) The children, who had eaten their fill, were allowed to leave the table.

Here the adjectival clause provides the reason for the occurrence of the main verb event. Such clauses might be excluded by the second criterion from the category of supplements due to their inability to move to a position other than post-subject; however, the equivalent *-ing* clause, *having eaten their fill*, can be fronted to pre-subject position, as in (7):

(7) Having eaten their fill, the children were allowed to leave the table.

This makes the inability of the non-restrictive adjectival clause in (6) to move to a non-post-subject position appear attributable to the need for the antecedent of the relative pronoun to occur before the pronoun itself, which has nothing to do with supplemental status.

Recourse to the criterion of omissibility is also fraught with problems. Following De Smet (2015), Bouzada-Jabois analyzes the *-ing* clauses in (8) and (9) below as "optional and therefore supplemental" (p. 87):

(8) At night workers just sat around *playing cards or sleeping*.

(9) (...) merchants who stood by the door of the custom-house *watching the disembarkation of a cargo*.

On the methodological level, treating these two clauses as 'optional' implies a view of the sentences containing them as abstract sequences detached from the intentions of the speaker/writer who produces them. In no way are the *-ing* clauses in (8) and (9) optional with respect to the speaker's intended message, however. The optionality test simply shows that the circumstantial adverb *around* and the circumstantial prepositional phrase *by the door of the custom-house* define the verbs *sit* and *stand* sufficiently for them to

make sense as predicates without the subsequent *-ing* clauses. It is very risky to draw conclusions about the structure of these two sentences based on such a criterion. Moreover, if one applies the criterion that the author borrows from De Smet (2015) for distinguishing the supplement clauses in (8) and (9) from the complement integrated participial clause in (10) below, according to which the complement can be identified by the fact that its omission "broadens the semantic scope of the main clause" (p. 89), the two purported supplements would also qualify as complements:

(10) The receptionist is busy filling a fifth box.

Just as *The receptionist is busy* has a broader semantic scope than the verbal predicate in (10) above, so the truncated predicates in *Workers just sat around* and *Merchants stood by the door of the custom-house* have a broader semantic scope than the full ones in (8) and (9).

The author subscribes to De Smet (2015)'s conclusion that the reason for the obligatoriness of the participle clause in the *spend time* construction illustrated in (11) below is "pragmatic rather than syntactic" (p. 91):

(11) (…) and she spent the entire evening convincing her that Uts was desperately passionately in love with her.

This claim is purported to be supported by the fact that the participle clause may be omitted if the time-word carries extra modification, as in (12), or is followed by a prepositional phrase or adverbial, as in (13):

(12) Julie spent a *restless and weary* evening, which passed into a restless and weary night

(13) She arrived in Jamaica in April, intending to spend six months *there*.

The purportedly pragmatic character of the obligatoriness of *convincing her that Uts was desperately passionately in love with her* leads Bouzada-Jabois to exclude such constructions from her corpus. One may legitimately question however whether the presence of a prepositional phrase, such as *with Susan* in (14) below, which would be included in Bouzada-Jabois' corpus due to the acceptability of *she spent the entire evening with Susan*, fundamentally changes the role of the participle clause in the construction instantiated by (11) above:

(14) (…) and she spent the entire evening *with Susan* convincing her that Uts was desperately passionately in love with her.

On a more general level, there are fundamental problems with the distinction adopted by the author between 'syntax' (complementation defined as the determination of arguments by a predicate) and 'pragmatics' (obligatoriness of certain adjuncts due to discourse requirements). Goldberg and Ackerman (2001) propose that obligatory adjuncts such as those occurring with the passives of accomplishment verbs (*This house was built last year* versus **This house was built*) can be accounted for by pragmatic requirements, in this case the need for the utterance to have an informational focus. Thus, *This house was built* does not provide significant information about the house, since we know that all houses are built. This observation raises the very important question of the contribution made to the determination of obligatoriness by pragmatic factors, which obviously have nothing to do with clause structure.[1] The idea behind the complement/adjunct distinction is that a complement is required in order to complete the meaning of its head, without which the latter would be incoherent, while an adjunct merely adds a further characterization to its head, restricting the latter to a proper subset of its denotation (see Dowty 2003: 34). However, it is questionable whether one can determine essentialness versus accidentalness outside of a context: thus, for example, the verb *tell* is usually treated as a three-place predicate involving an agent, a patient and an addressee; however, in a use such as (15) below there are but two arguments and there is no feeling at all that the other one has been ellipted:

(15) The author tells the story using a third person.

Some authors hold therefore that no diagnostic criteria have emerged that will reliably distinguish adjuncts from complements, e.g., Dowty (2003) or Herbst (2020). This undermines the syntax vs. pragmatics distinction at the basis of Bouzada-Jabois' delimitation of her corpus (p. 94), according to which

> all of the constructions included in this analysis may be regarded as completely optional elements because they are not syntactically required by the main clause in any sense and therefore do not take part in the complementation pattern of the main verb.

In the chapter on supplements in Late Modern English, the author examines the formal (mainly positional) and semantic features of these constructions. In the section on semantic features, she employs Kortmann's (1991: 121) scale of informativeness in order to classify the 19 adverbial meanings found in the corpus into four broad

[1] The fact that one could accept *This house was built* in a fairy-tale, as in *This house was built, but that one just appeared out of nowhere*, confirms the importance of pragmatic considerations for this question.

categories: 1) CCCC+ (which includes concession, contrast, condition, purpose, cause, result and concessive-conditional meanings), 2) temporal (which includes anteriority, posteriority and simultaneity), 3) manner, and 4) elaboration (which includes accompanying circumstances, addition, specification, exemplification, comparison, substitution and deictic-representational supplements). Her adaptation of Kortmann's scale raises a couple of problems. Firstly, Bouzada-Jabois does not follow the scale for the ranking of 'simultaneity', which is classified as less informative than 'manner' by Kortmann, nor for 'specification/exemplification', which are classified as more informative than 'simultaneity' in Kortmann's analysis. This departure from the original scale is neither mentioned nor justified. The second problem is that Kortmann's (1991: 120) scale was constructed exclusively for "present-participial free adjuncts/absolutes" and Bouzada-Jabois makes no adjustment for the *-ed* participles which are part of her corpus data. This is a critical defect for the temporal readings, as Kortmann justifies placing 'simultaneity' very low on the informativeness scale because he assumes it to be the unmarked value for the present participle. The unmarked value for the past participle, however, would not be 'simultaneity' but 'anteriority'.

Unresolved issues also arise in the discussion of the augmentation of supplements by means of connectors such as *with*, *rather than*, *besides*, *while*, etc. The received wisdom regarding the presence of connectors (see Kortmann 1991; Fonteyn and van de Pol 2016) holds that the more informative the meaning of a supplement, the more likely it is to be marked by a connector. However, the number one adverbial meaning marked by a connector in Bouzada-Jabois' corpus —'manner'— is located in the lower half of Kortmann's scale of informativeness and, in addition, the lowest member of Kortmann's scale —'accompanying circumstance'— ranks near the top of the list of adverbial meanings signaled by a connector[2] in Bouzada-Jabois' data. The author gives two reasons why 'manner' is thus ranked (p. 238). The first is that the manner category contains a great number of *-ing* forms that are introduced by the preposition *by* which could be claimed to be gerundive and so nominal rather than verbal. This argument does not carry much weight, however, as Bouzada-Jabois herself argues against it (pp. 39–40), demonstrating that such forms are verbal and not nominal. The second reason adduced is that Fonteyn and van de Pol (2016) regard 'manner' as one of the most informative adverbial categories. Since this stands in direct contradiction to Kortmann's

[2] This is also the case for the Present-day English data (pp. 306–307), although accompanying circumstance is the sixth rather than fourth among the most frequently augmented adverbial supplement.

scale, one would have expected some discussion of the superiority of Fonteyn and van de Pol's claim. Disappointingly, none is provided. Concerning the other problematic category, 'accompanying circumstance', Bouzada-Jabois observes that the augmented occurrence of this type represents only 21 percent of the total occurrences of adverbials denoting accompanying circumstances, which makes non-augmentation the norm for this type of adverbial. That is indeed the case, but it does not explain the disconnect between informativeness and augmentation with this category. Moreover, Bouzada-Jabois fails to point out that three other categories that rank very high on the informativeness scale are majoritarily non-augmented, as only 13 percent of adverbials expressing cause, 10 percent of those expressing purpose and 0 percent of those expressing result are preceded by an augmentor.

As a final note, it could be pointed out that the evidence is even stronger than Bouzada-Jabois makes it out to be for her claim that the data indicate a marked crystallization of the status of supplements and absolute constructions as sentential peripheral elements in modern and contemporary English (p. 320). Not only does the data show a statistically significant decrease in the most informative types of supplements and absolutes from Late Modern English to Present-Day English but, overall, the frequency of supplements has declined by a whopping 70 percent over this period (as Bouzada-Jabois shows in the graph on p. 261) and that of absolute constructions by 12 percent (as shown by van de Pol and Cuyckens 2014). This finding thus represents a significant contribution to the study of the periphery of the sentence in the recent history of English.

REFERENCES

De Smet, Hendrik. 2015. Participle clauses between adverbial and complement. *Word* 61/1: 39–74.

Dowty, David. 2003. The dual analysis of adjuncts and complements in Categorial Grammar. In Ewald Lang, Claudia Maienborn and Cathrine Fabricius-Hansen eds. *Modifying Adjuncts*. Berlin: Mouton de Gruyter, 33–66.

Fonteyn, Lauren and Nikki van de Pol. 2016. Divide and conquer: The formation and functional dynamics of the modern English *-ing* clause network. *English Language and Linguistics* 20/2: 185–219.

Goldberg, Adele E. and Farrell Ackerman. 2001. The pragmatics of obligatory adjuncts. *Language* 77/4: 798–814.

Herbst, Thomas. 2020. Dependency and valency approaches. In Bas Aarts, Gill Bowie and Gergana Popovo eds. *The Oxford Handbook of English Grammar*. Oxford: Oxford University Press, 124–152.

Kortmann, Bernd. 1991. *Free Adjuncts and Absolutes in English: Problems of Control and Interpretation*. London: Routledge.

van de Pol, Nikki and Hubert Cuyckens. 2014. The diffusion of English absolutes: A diachronic register study. In Kristin Davidse, Caroline Gentens, Lobke Ghesquière and Lieven Vandelotte eds. *Corpus Interrogation and Grammatical Patterns*. Amsterdam: John Benjamins, 265–294.

Reviewed by
Patrick Duffley
Laval University
Department of Languages, Linguistics and Translation
1030 Ave des Sciences-Humaines
G1V 0A6. Quebec
Canada
e-mail: patrick.duffley@lli.ulaval.ca

RiCL Research in Corpus Linguistics

Review of Lastres-López, Cristina. 2021. *From Subordination to Insubordination: A Functional-pragmatic Approach to If/si-constructions in English, French and Spanish Spoken Discourse.* Bern: Peter Lang. ISBN: 978-3-034-34220-9. DOI: https://doi.org/10.3726/b18393

An Van linden
University of Liège and KU Leuven / Belgium

Lastres-López's (2021) monograph presents a corpus-based contrastive study of conditional constructions used in spoken discourse in English, French and Spanish. It adopts a semasiological perspective, focusing on clauses introduced by *if* or *si* ('if' in French and Spanish), and takes a functional-pragmatic approach. Based on a detailed study of 3,558 *if/si*-constructions, it proposes classifications of both full-fledged conditional constructions (consisting of a protasis and an apodosis) and insubordinate conditional constructions (*viz.* subclauses without an accompanying main clause, see Evans 2007), and reflects on the diachronic relation between these two types. The book consists of 6 chapters, each of which I will discuss in turn.

Chapter 1 presents a brief introduction to the study. It delineates the object of investigation, namely structures introduced by *if* in English and *si* in French and Spanish spoken discourse in contexts of subordination and insubordination; constructions where these conjunctions introduce indirect polar questions are excluded from analysis. It contextualizes the study by indicating how it fills gaps in the existing literature on the topic. In doing so, however, Lastres-López merely posits claims about studies on related topics being abundant or scarce; she fails to cite references in support (she only does so in Chapter 2). The chapter concludes with a brief outline of the book and with a short presentation of the research questions that will be tackled.

Chapter 2 sketches the theoretical background to the study. Its first part presents a literature review of earlier work on conditionals on the one hand, and of previous research

on insubordination on the other. The former is an adequate synthesis that starts with classical approaches to conditionals, based on degrees of hypotheticality and linked up with tense and mood patterns, and works towards research on conditionals from a functional-pragmatic perspective. Lastres-López thus arrives at a fairly comprehensive classification, in which she carefully points to correspondences between proposals by distinct authors. Incidentally, the distinction between predictive and non-predictive conditionals central to Dancygier (1993, 1998) is missing. Consistent with the set-up of her monograph, Lastres-López also presents earlier work on conditionals from a contrastive and a corpus-based perspective. She hence waits until Chapter 2 to motivate her claims made in Chapter 1 about her study filling gaps in the literature. Her discussion of research on insubordination introduces the phenomenon adequately and presents various proposals about the diachrony of insubordinate structures across languages, meticulously laying out how these relate to each other. Lastres-López then homes in on previous work on insubordination in English, French and Spanish. With respect to English, I was struck by the omission of D'Hertefelt's (2018) classification of conditional insubordination. The latter's work is rightly mentioned in the context of the distinction between insubordination and dependency shift, but discussion of D'Hertefelt's taxonomy of conditional insubordination arrived at for English (and other Germanic languages) is starkly absent from this monograph, while it did receive attention in Lastres-López (2018: 46-47), that is, D'Hertefelt's (2015) dissertation, reworked into D'Hertefelt (2018). The second part of Chapter 2, in turn, presents the theoretical framework adopted in the monograph, which is couched in Hallidayan thought. Lastres-López's classification of full-fledged conditionals into ideational, interpersonal, and textual ones is convincing, including her critical appraisal of Kaltenböck's (2016) work. However, she fails to suggest how this Hallidayan framework would apply to insubordinate structures and brings the chapter to an abrupt end.

Chapter 3 presents the methodological background to the corpus-based study. Lastres-López starts off by justifying her choice for comparable corpora rather than parallel (or translation) corpora and compares, on the basis of Biber's (1988) multi-dimensional model of register analysis, the two spoken registers selected: conversation and parliamentary discourse. Although the selected registers are generally considered to occupy opposite ends on the formal-informal continuum, she concludes that they differ along only two out of five dimensions in Biber's (1988) model, namely with respect to

'involved versus informational production' on the one hand, and 'explicit versus situation-dependent reference' on the other. The chapter then details the corpora chosen and the data retrieval process, including screenshots of the corpus interfaces used. From the three corpora of parliamentary discourse selected, Lastres-López extracted random 500-hit samples for the period 2000–2010, targeting the conditional conjunction *if/si*. The same queries were used for the corpora with the selected conversational data, from which exhaustive samples were retrieved of no more than 940 hits per language. Although the description is detailed enough to ensure replicability, the reader gets no information about the overall word count of the Spanish corpus of parliamentary discourse used, nor of the 2000–2010 selected time frame, for any language. For the conversational data, we do not get to know the size of the sub-corpora consulted for French and Spanish (the monologue data still need to be subtracted from the totals given in Table 5 on p. 66). It would have been nice if the chapter had concluded with a table summarizing the various samples extracted for the studies reported on in Chapter 4.

Chapter 4 indeed presents the 'meat' of the monograph, arranged into three contrastive case-studies. The first two concern conditional subordination and differ in terms of register, with the first case-study concentrating on parliamentary discourse and the second focusing on face-to-face informal conversation. They are organized in the same way, which adds consistency to the volume and allows for an interesting cross-register comparison (in Section 4.1.3). Both case-studies start off with an overview of the types of structures introduced by *if/si* in the corpus data, illustrating also the discarded cases.[1] Then the data are analyzed for the same five analytical parameters, namely the Hallidayan metafunction of the conditional, the degree of likelihood of the conditional, the position of the protasis, the markedness of the apodosis (with a linking device like *then*) and the modal auxiliary in the apodosis. For parliamentary discourse, Lastres-López finds that ideational conditionals prevail in all three languages, although Spanish stands out in showing significantly larger portions of interpersonal conditionals than English and French. Another interesting cross-linguistic difference is that, when the metafunction of the construction is cross-classified with degree of likelihood, French and Spanish interpersonal conditionals are predominantly real conditionals, while the English ones

[1] In explaining the small share of *if*-complement clauses in English parliamentary discourse (0.60%) compared to French (14.40%) and Spanish (21.80%), Lastres-López overlooks the availability of a contender for introducing indirect polar questions in English, namely *whether*, which French and Spanish lack (pp. 70–71).

show more variation between real and potential conditionals. A last cross-linguistic difference observed is not surprising: English shows a much higher ratio of modal verbs in the apodosis (67.45 %) than French (21.09%) and Spanish (12.78%) as, in the latter languages, meanings equivalent to *will* and *would* are coded by verbal endings (mood-tense combinations) on the finite lexical verb (p. 91). Unfortunately, the modals found in English have not been classified into semantic subtypes (e.g., epistemic, deontic, and dynamic modality), nor have the attested Romance auxiliaries. Such an analysis would have allowed more fine-grained conclusions with respect to this last parameter.

The second case-study has the same set-up as the first one, but involves two additional analytical parameters, both pertaining to interpersonal conditionals, which chalk up much higher shares in the conversational data studied than in parliamentary discourse. Again, relevant corpus hits are separated from irrelevant ones, and the latter are categorized into several subtypes and aptly illustrated with examples for each language. Here, while I see why repetitions and false starts are excluded from further analysis, I do not understand why discontinuous conditionals are (see note 41 on p. 95). Although these are either co-constructed or interrupted, as is not unexpected in spontaneous conversation, they nevertheless form complete conditionals.

The interpersonal conditionals receive special treatment. Firstly, they are further analyzed for an interpersonal subfunction according to Warchal's (2010) classification (epistemic, opinion/evaluation, politeness, relevance, reservation, and metalinguistic); Lastres-López here takes great care in explaining how the examples proffered instantiate the interpersonal subfunction in question. However, chances are missed to establish links with the literature review in Chapter 2. For instance, conditionals serving the politeness subfunction (pp. 105–106) could have been categorized as speech act conditionals as defined on p. 30, where the protasis in example (20) indeed mitigates the speaker's evaluation expressed in the apodosis. Likewise, relevance conditionals could have been classified as speech act conditionals as well: in example (126) on p. 107, the protasis justifies why the speaker utters the statement in the apodosis. Secondly, interpersonal conditionals are additionally coded for whether they convey stance (and are hence speaker-oriented) or engagement (and are hence addressee-oriented). Interestingly, Lastres-López cross-classifies this parameter with that of interpersonal subfunction, which shows that these are truly independent parameters. Even more interesting are the correlations revealed between the metafunction of the conditional and the position of the

protasis, and —among interpersonal conditionals— the correlations between the interpersonal subfunction of the conditional and again the position of the protasis. Across the three metafunctions, the protasis occurs predominantly in sentence-initial position across the three languages. Homing in on interpersonal conditionals, it becomes clear that this preference is mainly due to what is observed in epistemic conditionals. Unlike the latter, relevance, reservation, and metalinguistic conditionals prefer non-initial protases in English and French.[2] With respect to markedness of the apodosis, almost absent in parliamentary discourse, the detailed analysis of interpersonal conditionals in the conversational data uncovers noteworthy correlations. That is, marked apodoses are restricted to (interpersonal) epistemic conditionals in English and French. In Spanish, by contrast, this use accounts for only 54.55 percent of the marked apodoses in conversation, "with the remaining proportion distributed across a wide range of metafunctions and subfunctions" (p. 128). For the last parameter of modal auxiliaries in the apodosis, the conversational data show overall lower shares of modals than in parliamentary discourse. Regrettably, the conversational data were not further analyzed for semantic subtype of modal meaning either.

These first two case-studies reveal very interesting results, and hence significantly add to our knowledge about conditionals in English, French and Spanish, as set out above, but they also show some shortcomings, some of which I have already mentioned in the above paragraphs. First, I beg to disagree with Lastres-López's analysis of the corpus examples in (1) and (2) further below.

(1) So, if we want to increase the current 3,000 adoptees by at least 50 per cent, as we all do, there is plenty of scope in the existing material, and we need to concentrate on why more such people are not coming forward or being approved as adopters (Hansard Corpus – British Parliament) (Ex. (72), p. 78).

Example (1) is categorized as an ideational conditional, in spite of the comment that the conditional is used to render the message less assertive and that in similar examples "*if* can be paraphrased by *since*" (p. 78). These are exactly two features of what Dancygier (1993, 1998) calls non-predictive conditionals, that is, conditionals that lack a causal relation between protasis and apodosis, under which Dancygier (1993: 422–424) subsumes both Sweetser's (1990) epistemic and speech act conditionals, which Lastres-López in turn correctly classifies as interpersonal conditionals. In my view, in (1) the

[2] Spanish is the odd one out in showing a preference for sentence-initial protases in relevance conditionals (66.67 %), and in showing no reservation or metalinguistic conditionals at all (p. 119, note 43).

protasis expresses an assumption that is manifest to both speaker and hearer ('if it is really the case that we want to …') on the basis of which the speaker arrives at an inference with a deontic flavor in the apodosis, which pragmatically serves as a call for action ('let's concentrate on ...'). To me, then, (1) is an interpersonal conditional rather than an ideational one. The reverse goes for example (2).

(2) If you are born in the Gorbals and there's absolutely no chance of your having money well then you grow up as a normal Gorbals-born person (ICE-GB: S1A – 075 #090: 1: B) (Ex. (117), p. 104).

Example (2) is analyzed as an interpersonal conditional, more specifically an epistemic conditional, which can be "paraphrased as 'If *I assume* [protasis], then *I conclude* [apodosis]'" (p. 104). However, without further co-text, I would analyze (2) as a predictive conditional, that is, a conditional in which the protasis expresses an assumption on the basis of which the speaker arrives at a prediction in the apodosis (Dancygier 1993: 405–406). To my mind, there is a sequential and causal relation between the protasis and apodosis in (2), and the example hence serves the ideational metafunction rather than the interpersonal one. Needless to say, my reservations about Lastres-López's analyses of (1) and (2) impinge on my appraisal of these first two case-studies.

A second weakness relates to the absence of the notion of 'backshift', and the way it interacts with the metafunction of conditionals. Again, I turn to Dancygier (1993: 405–406) here, who shows that, in English, in predictive conditionals the interpretation of verb forms involves back-shifting: "the time reference intended by the speaker is systematically *later* than the time referred to by the verb form in its prototypical (non-conditional) uses" (emphasis original). This should have been discussed in the sections on degrees of likelihood of the two case-studies. Dancygier's (1993) observation that there is no back-shift in epistemic and speech act conditionals in English raises questions about the potential and unreal conditions serving the interpersonal metafunction in the English datasets of the two case-studies. In Dancygier's (1993: 417) terms:

> the verb forms in non-predictive conditionals refer to the time they indicate. In other words, they are not backshifted and can be used according to the rules governing non-conditional constructions.

Incidentally, Lastres-López restricts unreal conditions to "past time event(s) which cannot be changed" (p. 82), and hence seems to overlook the class of unreal conditions with

present-time reference, such as “counteridentical-P conditionals” (e.g., *If I were you*), described by Declerck and Reed (2001: 100).

I now turn to the third case-study in chapter 4, which focuses on conditional insubordination in the conversational data studied. Interestingly, it reveals stark cross-linguistic differences. For one, Spanish shows a much larger portion of insubordinate conditionals (20.85%) than English (4.18%) and French (2.24%) (p. 131). A second difference pertains to the type of discourse function served: whereas English insubordinate *if*-clauses mainly serve directive functions (requests, suggestions and offers), their French and Spanish counterparts are predominantly used to express assertions and exclamations. Unfortunately, adding to the fact that the typology of directive subfunctions presented in Table 28 (p. 133) lacks parameters that together uniquely define the five types distinguished, the case-study does not go in much detail regarding the results mentioned above. For instance, if the presence of modal auxiliaries is mentioned at all (e.g., for requests, but not for offers), there is no discussion of their semantic subtype, and no attention is given to back-shift or tense-mood marking of finite verbs more generally, or to polarity reversal in examples like (182) on p. 142. Incidentally, I wonder whether the prosodic mark-up in example (179) on p. 141 does not suggest that French *si* functions as a positive polarity item here rather than a conditional conjunction. Also, I am in doubt as to whether the *si*-clause in (181) on p. 142 is not a postposed epistemic conditional rather than an insubordinate one: that is, I would accept an analysis of (181) as a bridging context supporting both an interpersonal (epistemic) conditional reading and an insubordinate reading.

For the three case-studies, Lastres-López nicely combines qualitative and quantitative analyses, and the figures in the latter always add up. She also presents the results of statistical tests in graphs, plotting 95 percent Wilson confidence intervals, but here I was often confused as what these graphs do and do not show. While it is stated in note 38 on p. 76 that “[w]hen the confidence intervals (in the form of I-shaped bars) do not overlap at any point, the results are statistically significant,” in multiple graphs the bars do not overlap but yet only some differences are said to be statistically significant and other (also without overlap) are not, and the reader is supposed to see this in the respective graphs (e.g. for Figures 10, 17, and 21). I was puzzled by the discussion of these graphs.

Chapter 5, then, reflects on the developmental relations between the constructions studied, and feeds the synchronic findings of Chapter 4 into a diachronic hypothesis. Specifically, it puts forward a pathway of pragmaticalization, along which ideational conditionals acquire interpersonal and textual functions in full conditional constructions, which in turn develop into insubordinate constructions and pragmatic markers like *if you choose/ like/ prefer/ want/ wish*. Although this pathway is intuitively appealing and in line with numerous proposals posited for similar phenomena, it remains sterile in that Lastres-López does not specify which interpersonal subtypes would develop into which insubordinate subtypes. Nor does she point to bridging contexts to motivate the pathway and, hence, seems to underexploit her dataset (see my comment above relating to ex. (181) on p. 142).

Chapter 5 rounds off with a detailed summary, strangely marked for present tense, and Chapter 6 offers some avenues for further research.

Overall, Lastres-López uses an engaging writing style, and her monograph contains only a handful of typos or infelicities (e.g., *smallest* for *smaller* in "The smallest the confidence intervals, the greater the level of certainty on the observed values" on p. 76). However, in terms of local text organization, I often felt that examples were given too late. The long distance between the introduction of an example in the running text and the presentation of the example itself puts a strain on the reader, and often also affects indentation (I bet that in relation to the latter it is the publisher's typesetting rules that are to blame, not the author). At a higher level of text organization, I regret the use of sections that only have one subsection. For instance, no separate subsection had to be assigned to Section 4.1.2.2.1, as there is no Section 4.1.2.2.2 to differentiate it from. The level of Section 4.2.1 is likewise redundant, as there is no Section 4.2.2.

In conclusion, while there is certainly room for improvement, I think Lastres-López put together a very interesting monograph, substantially contributing to the domain of contrastive corpus linguistics and significantly advancing our understanding of conditionals, whether in full-fledged complex sentences or used independently, in English, French and Spanish spoken discourse.

References

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

Dancygier, Barbara. 1993. Interpreting conditionals: Time, knowledge, and causation. *Journal of Pragmatics* 19/5: 403–434.

Dancygier, Barbara. 1998. *Conditionals and Prediction: Time, Knowledge, and Causation in Conditional Constructions*. Cambridge: Cambridge University Press.

Declerck, Renaat and Susan Reed. 2001. *Conditionals: A Comprehensive Empirical Analysis*. Berlin: Mouton de Gruyter.

D'Hertefelt, Sarah, 2015. *Insubordination in Germanic: A Typology of Complement and Conditional Constructions*. Leuven: KU Leuven University Dissertation.

D'Hertefelt, Sarah. 2018. *Insubordination in Germanic: A Typology of Complement and Conditional Constructions*. Berlin: Mouton de Gruyter.

Evans, Nicholas. 2007. Insubordination and its uses. In Irina Nikolaeva ed. *Finiteness: Theoretical and Empirical Foundations*. Oxford: Oxford University Press, 366–431.

Kaltenböck, Gunther. 2016. On the grammatical status of insubordinate if-clauses. In Gunther Kaltenböck, Evelein Keizer and Arne Lohman eds. *Outside the Clause: Form and Function of Extra-clausal Constituents*. Amsterdam: John Benjamins, 341–377.

Lastres-López, Cristina. 2018. *If*-insubordination in spoken British English: Syntactic and pragmatic properties. *Language Sciences* 66: 42–59.

Sweetser, Eve. 1990. *From Etymology to Pragmatics*. Cambridge: Cambridge University Press.

Warchal, Krystyna. 2010. Moulding interpersonal relations through conditional clauses: Consensus-building strategies in written academic discourse. *Journal of English for Academic Purposes* 9/2: 140–150.

Reviewed by
An Van linden
University of Liège
Faculty of Philosophy and Arts
Department of Modern Languages: Linguistics, Literature and Translation
Place Cockerill 3–5
4000 Liège
Belgium
e-mail: an.vanlinden@uliege.be