

Research in Corpus Linguistics



RiCL 10/1 (2022)

Editors

Paula Rodríguez-Puente and Carlos Prado-Alonso

ISSN 2243-4712

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

RiCL

Research in
Corpus Linguistics



Official journal of

aelinco

Asociación Española de Lingüística de Corpus

Articles	Pages
Evaluative stance in Vietnamese and English writing by the same authors: A corpus-informed appraisal study Tieu-Thuy Chung, Luyen-Thi Bui, Peter Crosthwaite	1–30
The Process Corpus of English in Education: Going beyond the written text Gaëtanelle Gilquin	31–44
The compilation of a developmental spoken English corpus of Turkish EFL learners Ece Genç-Yöntem, Evrim Eveyik-Aydın	45–62
How is information content distributed in RA introductions across disciplines? An entropy-based approach Wei Xiao, Jin Liu, Li Li	63–83
Libya, the media and the language of violence: A Corpus-Assisted Discourse Analysis Safa Attia	84–116
The FGLOCTweet Corpus: An English tweet-based corpus for fine-grained location-detection tasks Nicolás José Fernández-Martínez	117–133
A corpus study of the term evidence in open peer reviews to research articles in the British Medical Journal Ingrid García-Ostbye, Barry Pennock-Speck	134–155
Preterit-imperfect acquisition in L2 Spanish writing: Moving beyond lexical aspect Sophia Minnillo, Claudia Sánchez-Gutiérrez, Agustina Carando, Samuel Davidson, Paloma Fernández Mira, Kenji Sagae	156–184
Book Reviews	
Review of Pérez-Paredes, Pascual. 2021. <i>Corpus Linguistics for Education: A Guide for Research</i>. London: Routledge. ISBN: 978-0-367-19843-5. DOI: https://doi.org/10.4324/9780429243615 Barry Pennock-Speck	185–191
Review of Bouso, Tamara. 2021. <i>Changes in Argument Structure: The Transitivity Reaction Object Construction</i>. Bern: Peter Lang. ISBN: 978-3-034-34095-3. https://doi.org/10.3726/b17960 Sune Gregersen	192–204
Review of Moskovich, Isabel, Inés Lareo and Gonzalo Camiña eds. “All Families and Genera” <i>Exploring the Corpus of English Life Sciences Texts</i>. Amsterdam: John Benjamins. ISBN: 978-9-027-20924-5. https://doi.org/10.1075/z.237 Stephanie Degaetano-Ortlieb	205–212
Review of Castro-Chao, Noelia. 2021. <i>Argument Structure in Flux: The Development of Impersonal Constructions in Middle and Early Modern English, with Special Reference to Verbs of Desire</i>. Bern: Peter Lang. ISBN: 978-3-034-34189-9. DOI: https://doi.org/10.3726/b17694 Ayumi Miura	213–227
Review of Wallis, Sean. 2020. <i>Statistics in Corpus Linguistics: A New Approach</i>. London: Routledge. ISBN: 978-1-138-58938-4. DOI: https://doi.org/10.4324/9780429491696 Tove Larsson	228–232

Evaluative stance in Vietnamese and English writing by the same authors: A corpus-informed *appraisal* study

Tieu-Thuy Chung^a – Luyen-Thi Bui^a – Peter Crosthwaite^b
Tra Vinh University^a / Vietnam
University of Queensland^b / Australia

Abstract – *Appraisal* theory (Martin and White 2005), an approach to discourse analysis dealing with evaluative language, has been previously employed in analysing newspaper articles and spoken discourses in several earlier studies, although it is gaining in popularity as a framework for comparing first and second (L1/L2) writing. This study investigated 40 English majors' Vietnamese and English paragraphs for evaluative language, a key component of successful academic writing, as realised under *Appraisal* theory. To this purpose, we collected L1 Vietnamese and L2 English data from the same student writers across the same topics and using a corpus-informed Contrastive Interlanguage Analysis approach to the annotation and analysis of *appraisal*. A range of commonalities were present in the use of *appraisal* across the two language varieties, while the results also suggest significant differences between students' evaluative expressions in Vietnamese as a mother tongue and English as a second or foreign language. This variation includes the comparative under- and over-use of specific *appraisal* resources employed in L1 and L2 writing respectively, in particular, regarding writers' employment of attitudinal features. The findings serve to inform future pedagogical applications regarding explicit instruction in stance and *appraisal* features for novice L2 English writers in Vietnam.

Keywords – L2 writing; Vietnamese; corpora; evaluation; *Appraisal* theory; stance

1. BACKGROUND

In Vietnam, English has increased in prominence as the main foreign language taught in primary, secondary and tertiary institutions. As many second language (L2) learners in other countries, Vietnamese students do not use English for everyday communication outside the classroom, so it does not seem easy for them to master L2 English. Of the four skills, writing has been found to be one of the most difficult for Vietnamese learners to acquire in both Vietnamese as a first language (L1), as well as L2 English, as observed in Bailey (2006: vii) in that such students “often find the written demands of their courses very challenging.” For this reason, there is now an increasing amount of research dealing with L2 English writing in both Vietnam and other nearby countries, such as Cambodia,

with studies using various methods across a range of linguistic perspectives on L1 and L2 writing for eventual use by English language teachers, educators, and learners.

This trend is also seen in recent studies on evaluative language, a key component of argumentation in academic writing. Hunston and Thompson's (2000) definition of evaluation (as cited in Lam and Crosthwaite 2018: 9) is

the expression of the speaker or writer's attitude or stance towards, viewpoint on, or feelings about the entities or propositions that he or she is talking about.

There have been different definitions and terms associated with evaluation, with a number of frameworks present in the literature (e.g. Goffman 1981; Labov 1984; Chafe and Nichols 1986; Biber and Finegan 1989; Ochs 1989; Simpson 1993; Hyland 2005). A number of recent studies have adopted Martin and White's (2005) comprehensive *appraisal* framework to focus on the interpersonal aspect of language, although there has been little research using this framework to study L1/L2 writing in South-East Asian contexts, a gap this study intends to fill.

Employing *appraisal* theory as a research tool to analyse students' writing, preceding researchers have drawn important conclusions. However, there is still a lack of *appraisal* studies where the same L2 English and L1 Vietnamese writing tasks were conducted by the same writers, which could allow researchers to more accurately determine the nature of L1 transfer on L2 employment of evaluative language in future studies. Seeing the potential for understanding this phenomenon, we seek to address the following research question: How do the same writers writing in both L1 Vietnamese and L2 English project evaluative stance as realised in terms of the *attitude*, *engagement* and *graduation* domains under *appraisal* theory?

In line with this investigation, this paper adopts a corpus-informed approach. According to McEnery and Hardie (2012: 17), the corpus-informed approach utilises "only selected parts of a corpus" and the corpus is considered "simply as a bank of examples to illustrate a theory." This approach was previously used in Lam and Crosthwaite (2018), who analysed all three domains of *appraisal* across texts in L1 English and L2 English written by L1 Cantonese speakers in Hong Kong, produced on the same tasks and under the same conditions, and finding significant variation between L1/L2 writers in the *appraisal* resources employed. We seek to replicate this approach in the current study.

2. THEORETICAL FRAMEWORK

Initiated from the project *Write it Right* and inspired by Halliday's systemic functional linguistic theory, Martin and White (2005) introduced *Appraisal* theory, dealing with the *interpersonal meaning of evaluative language* in written discourse.

It is concerned with how writers/speakers approve and disapprove, enthuse and abhor, applaud and criticise, and with how they position their readers/listeners to do likewise. It is concerned with the construction by texts of communities of shared feelings and values, and with the linguistic mechanisms for the sharing of emotions, tastes and normative assessments. It is concerned with how writers/speakers construe for themselves particular authorial identities or personae, with how they align or disalign themselves with actual or potential respondents, and with how they construct for their texts an intended or ideal audience. (Martin and White 2005: 1).

This theory includes the study of evaluative language over three domains: *attitude*, *engagement*, and *graduation*.

2.1. Attitude

Attitude has three subdomains: *appreciation*, *judgement*, and *affect*. *Affect* reveals “positive and negative feelings” (Martin and White 2005: 42), *judgement* shows admiration for, criticism and condemnation of behaviour, while the assessment of text, process, or natural phenomena belongs to *appreciation*. The interconnection between these subtypes is demonstrated in Figure 1.

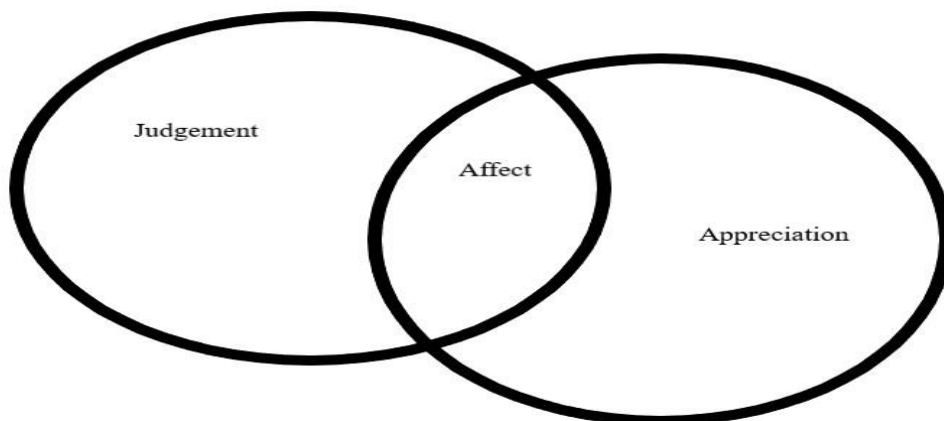


Figure 1: Interconnection between *attitude* domain adapted from Martin and White (2005: 45)

Affect is divided into four subcategories: *un/happiness*, *dis/satisfaction*, *in/security*, and *dis/inclination*. Examples (1) to (4) (from Martin and White 2005: 49, 51) illustrate these resources as follows.

(1) The captain felt **sad/ happy** [*un/happiness*].

(2) The captain felt **fed up/ absorbed** [*dis/satisfaction*].

(3) The captain felt **anxious/ confident** [*in/security*].

(4) Linda is **wary about/ longing for** her upcoming presentation [*dis/inclination*].

Judgement has two main subtypes, *social sanction* and *social esteem*, divided into five further subtypes including *normality*, *capacity*, *tenacity*, *propriety*, and *veracity*. *Normality* answers the question *How special?*, as in (5). Instances, such as (6), responding to the question *How capable?*, belong to *capacity*. *Tenacity* shows the answer for the question *How dependable?*, as in (7). *How honest?* questions of *veracity* as with (8). Finally, example (9), answering *How far beyond reproach?*, is an instance of *propriety* (Martin and White 2005: 53).

(5) I am **unlucky/lucky**.

(6) Mary is **immature/mature**.

(7) He is **timid/brave**.

(8) The woman is **dishonest/honest**.

(9) That captain is **immoral/moral**.

Appreciation consists of *reaction*, *composition*, and *social valuation* (Martin and White 2005: 56). While sentences such as that in (10) illustrate *reaction*, that in example (11) is an example of *composition*, while (12) relates to *social valuation*.

(10) The movie is **boring**.

(11) The young woman looks **shapely**.

(12) This writing is **original**.

Table 1 summarises the sub-categories of the *attitude* domain.

Attitude	Affect	<i>un/happiness</i>
		<i>dis/satisfaction</i>
		<i>in/security,</i>
	Judgement	<i>dis/inclination</i>
		<i>normality</i>
		<i>capacity</i>
		<i>tenacity</i>
		<i>propriety</i>
		<i>veracity</i>
	Appreciation	<i>reaction</i>
		<i>composition</i>
		<i>social valuation</i>

Table 1: The *attitude* domain adapted from Martin and White (2005: 45–58)

2.2. Engagement

Engagement evaluates whether the writer uses a single voice (*monoglossia*) or recognises dialogistic alternatives (*heteroglossia*) when expressing his or her ideas. The examples in (13) and (14) (from Martin and White 2005: 100) illustrate *engagement* resources:

(13) The banks have been greedy [*monoglossic*].

(14) **In my view** the banks have been greedy [*heteroglossic*].

In other words, *engagement* reveals if *bare assertions* or *expansive* and *contractive* options are employed by the writer.

Expansive options include *entertain* and *attribute*, while *contractive* ones consist of *disclaim* and *proclaim*. *Entertain* examples can be found in (15a–b), while (16a–b) are instances of *attribution*. *Disclamation* is divided into *negation*, as in (17) or *concession* as in (18). *Proclamation* has further subtypes including *concurrence* as in (19), *justification* as in (20), *pronouncement* as in (21), and *endorsement* as in (22) (Martin and White 2005: 100–127). More specific information can be found in Figure 2.

(15a) **I believe** he did this.

(15b) It **seems** that he did this.

(16a) **Some believe** that he did this.

(16b) **It is rumoured that** he did this.

(17) I **didn't** see her.

(18) It is raining, **but** I want to go out.

(19) **Naturally**, they enter the competition.

(20) He did this **because** he wanted to make me surprised.

(21) I **contend** that you have decided to join this.

(22) They **have shown** Mary did enter the room.

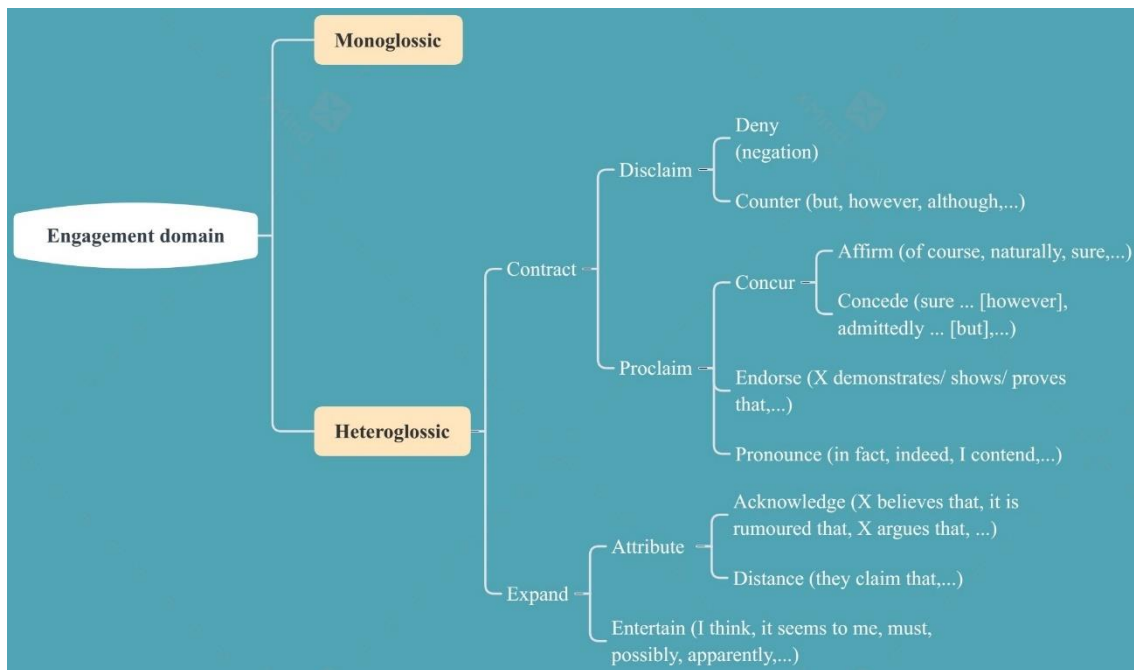


Figure 2: *Engagement domain* (adapted from Martin and White 2005: 134)

2.3. Graduation

Graduation deals with the notions of *force* and *focus*. *Force* (Martin and White 2005: 141–149) indicates the scalability of *intensification* and *quantification*. *Focus* (Martin and White 2005: 137) expresses the *sharpening* or *softening* of semantic boundaries. Examples are provided in (23) to (26), and further examples are illustrated in Figure 3.

(23) This film is **very** [*intensification*] interesting.

(24) They have made **many** [*quantification*] good friends.

(25) They don't play **real** jazz [*sharpen*].

(26) They play jazz, **sort of** [*soften*].

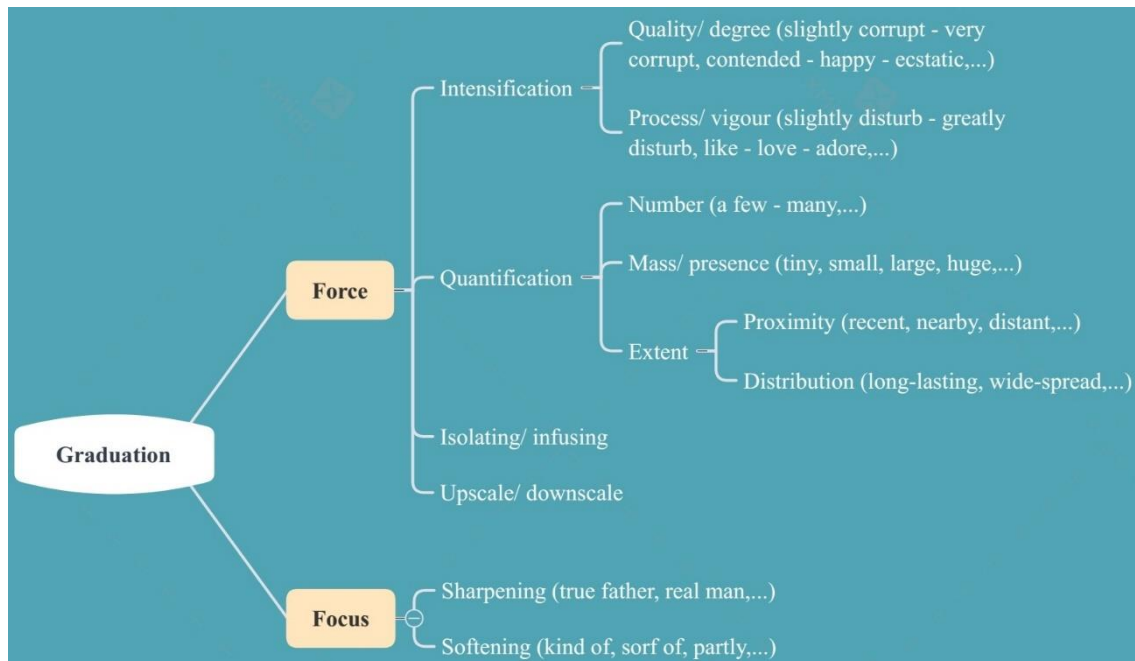


Figure 3: *Graduation* domain adapted from Martin and White (2005: 154)

3. PREVIOUS INVESTIGATIONS INTO APPRAISAL FOR L2 WRITING

Appraisal has been a topic of interest attracting a large body of research (e.g. Coffin and Hewings 2004; Hood 2004, 2006; Derewianka 2007; Swain 2007; Lancaster 2011; Geng and Wharton 2016, among others). Comparing the employment of *attitude* in research introductions written by L2 English learners in a tertiary education setting with experts' L1 writing, Hood (2004) recognised a combination of evaluative values was employed in student writers' texts. Explicit emotion and judgements on behaviour were adopted in evaluation in students' writing suggestive of a personalised treatment within their introductory passages, whereas experts' texts featured frequent *appreciation*. Geng and Wharton (2016) conducted a study on *engagement*, recognising that L2 English writers of L1 Mandarin background were affected by negative L1 transfer of *engagement* features

while trying to convey their stance in L2. This finding helps to explain why low-graded English essays in Coffin and Hewings (2004) tended to heavily feature *pronounce* features, or the expression of overtly authorial voice, resulting in making writers' claims less persuasive, as would be found in Mandarin. These two studies indicate that balancing dialogic expansion and contraction remains an issue for L2 writers. Lancaster (2011: 18) suggests that students who were more proficient in argumentative and critically reasoning writing tended "to be authoritative and dialogically open," as opposed to lower level learners. Exploring the adoption of *graduation* in L2 students' writing, both Hood (2006) and Derewianka (2007) agreed that L2 learners need not only be taught relevant linguistic devices but also how to manage such resources to enhance their production of evaluative values.

As mentioned, in Vietnam there have been several studies employing *Appraisal* theory in analysing newspaper articles (Vo 2011; Vo 2017) and spoken discourses (Tran 2011; Ngo and Unsworth 2015), although works on students' L2 English writing using this approach are still limited. Ho (2011) conducted a contrastive study comparing his students' L2 English essays with experts' L1 Vietnamese and L1 English essays, using the *engagement* domain of *appraisal* theory as his research tool. One of the findings is that L2 English Vietnamese students produced expanding resources in their English texts more frequently than those in L1 native and L1 English essays. However, students' over-reliance on their personal points of view may lead to a perceived lack of persuasiveness in their L2 writing. Chung (2018) found in her students' L2 intermediate English paragraphs the two subcategories of *attitude*, *affect* and *appreciation* were used in relative balance, while *judgement* was predominantly employed when making evaluation, and *heteroglossia* was adopted twice as much as *monoglossia*. L2 writers of Vietnamese background in this study were inclined to assess behaviour more than to express feelings or evaluate things, a distribution considered against the norms of academic writing.

4. METHODS

4.1. Corpus data

38 Vietnamese third-year tertiary students majoring in English were enrolled in a course named *Vietnamese in Practice* from February to June 2018 in Tra Vinh University in Vietnam. In this course, they were taught how to use L1 Vietnamese properly from word

choice to sentence and paragraph level (many students entering tertiary education lack experience in writing in L1). Paragraph writing was a requirement for both mid and final examinations, and therefore constitutes the unit of investigation for the present study. All procedures were performed in compliance with relevant laws and institutional guidelines and they have been approved by the appropriate institutional committee. Informed consent was obtained for experimentation with human subjects and the privacy rights of human subjects must always be observed. Participant information is shown in Table 2.

No.	Language Group	Quantity	L2 English Proficiency	Notes
1	Vietnamese	38	Intermediate	Females – 25
2	English	38		Males – 13 Ages – 18–21 = 38

Table 2: Course participants

Students were asked to write in Vietnamese first and in English later with a one-week time gap. Both tests were limited in timeframe and under the supervision of the lecturer. The first writing topic was based on prior reading about the meaning of *narcissus*, and students were asked to write their own understanding of this term in a short paragraph. The second topic asked students to reflect on a certain lifestyle from an excerpt they had read in a previous question. The selected data for analysis were chosen randomly from the entire pool of students' writing regardless of gender or age (Table 3). While small, the corpus is suitable for a corpus-informed (rather than corpus-based) approach as seen in other *appraisal* studies such as Lam and Crosthwaite (2018).

Task	L2 English		L1 Vietnamese	
	Texts	Words	Texts	Words
Flower	10	521	10	855
Lifestyle	10	963	10	1,436
Total	20	1,484	20	2,291
Total corpus size = (n=40)				

Table 3: Corpus detail

4.2. Research instruments and coding

We employed a corpus-informed approach following a Contrastive Interlanguage Analysis methodology (CIA²) proposed by Granger (2015) in terms of the text types and language varieties under investigation (Figure 4). Granger (1996: 43) states “CIA does not established comparison between two different languages but between native and learner varieties of one and the same language.” However, as a response to accusations

of the comparative fallacy (Bley-Vroman 1983), Granger (2015) introduces CIA² in which the Interlanguage Varieties (ILV) referring to the learner language can be the learner's mother tongue.

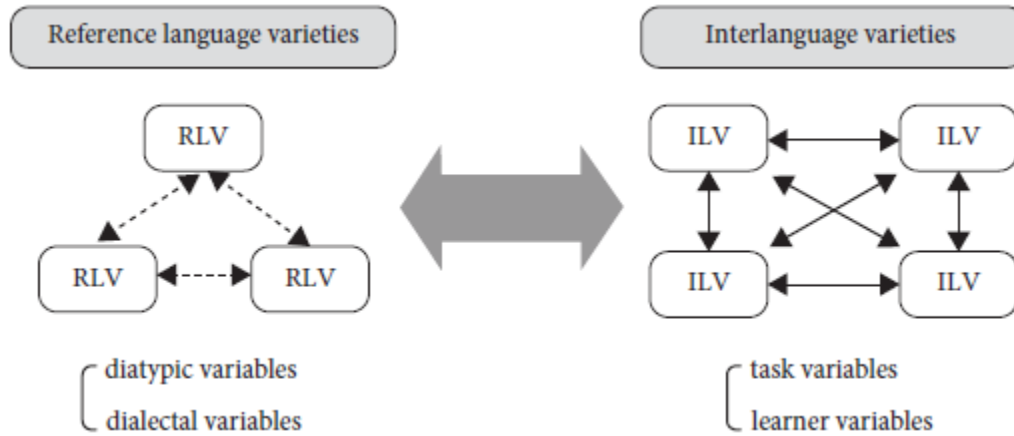


Figure 4: CIA² sourced (from Granger 2015: 17)

Diatypic variables may consist of register (field, mode, tenor) or text types. In addition, dialectal variables can include (non-)standard dialects, regional dialects, social dialects, temporal dialects, or expert/novice dialects. The task variables might range from complexity to genres while the learner varieties can be their L2, third language (L3), or their mother tongue.

The paragraphs were annotated following *Appraisal* theory, including the components *engagement*, *attitude*, and *graduation*, using *UAMCorpusTool* (v3.3) software developed by O'Donnell (2016). This study combined a qualitative method in identifying the similar and different *appraisal* resources used in students' L1 Vietnamese and L2 English writing, and quantitative methods in calculating the frequency of coded features by UAM software. Since the original *appraisal* framework may be considered overly comprehensive for our purposes, this paper adapts Martin and White (2005) as well as Lam and Crosthwaite (2018) to build a simplified version of the coding scheme (see Figure 5). The three primary domains are kept with a reduced set of subcategories per domain. Specifically, besides *affect*, *judgement* and *appreciation*, the *attitude* domain needs to consider (non-)authorial evaluation of emotions, explicit/implicit evaluation and the valence of attitudinal resources. The multi-voiced argumentation of *engagement* is explored through contraction (denials, countering, concurring, endorsing, justifying, and pronouncing) and expansion (entertaining and attributing). Meanwhile, the *graduation*

domain features the scaling of *force*, while the vagueness or exactness of attitudinal values are managed through *focus*.

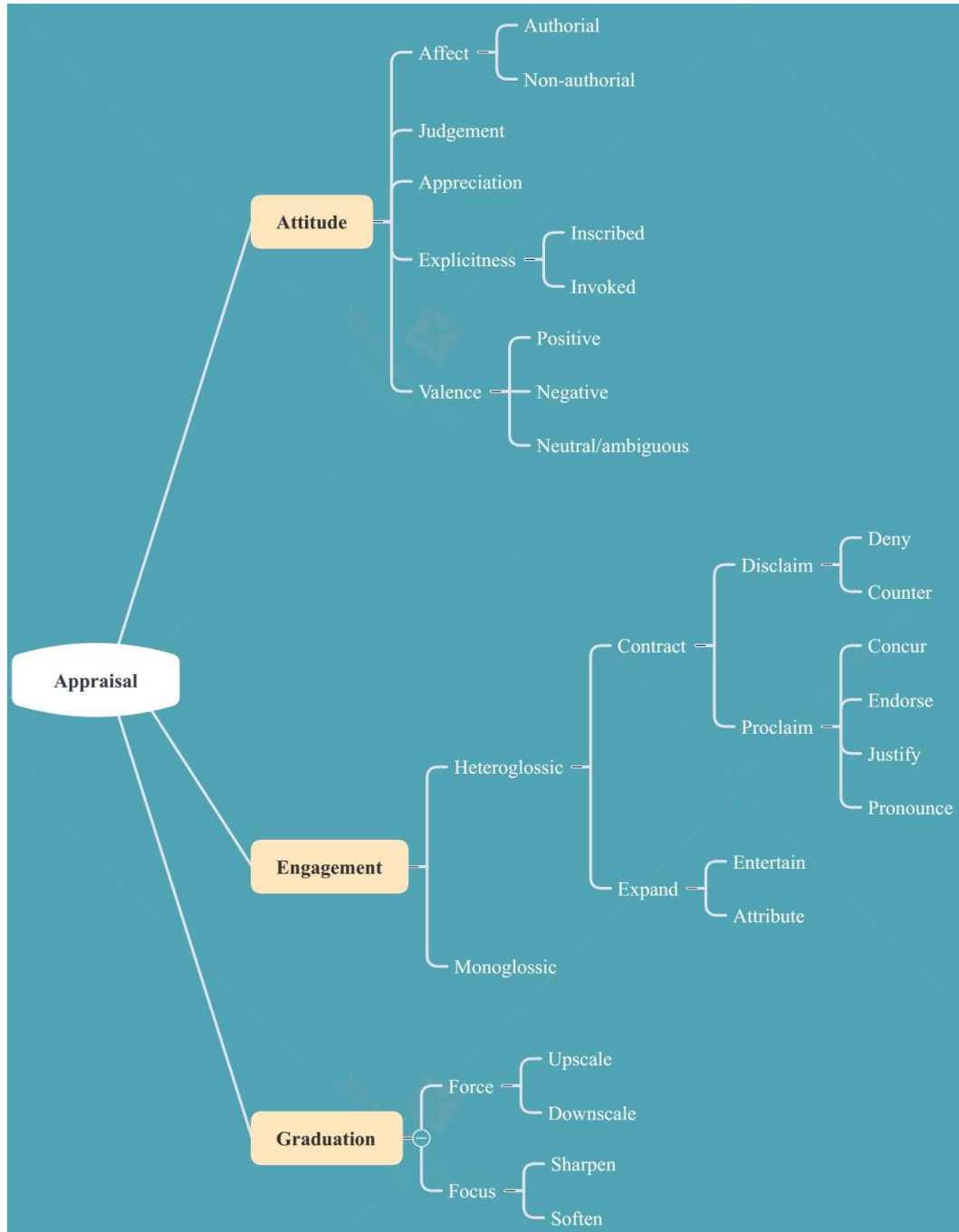


Figure 5: The simplified *appraisal* framework (adapted from Martin and White 2005: 38; Lam and Crosthwaite 2018: 20)

The coding was repeated twice at the interval of two weeks to help the coder gain better understanding and improve the validity of this process. In line with Lam and Crosthwaite (2018), the Intraclass Correlation Coefficient (ICC), a statistical measure of rater agreement, was employed to measure the stability of coding. We implemented ICC using

a two-way random model to assess the intra-coder agreement. The ICC result was over .88 (Table 4), which authenticates ‘good’ reliability (Koo and Li 2016).

		ICC results		
Reliability		<i>Attitude</i>	<i>Engagement</i>	<i>Graduation</i>
Stability	Single measures	.938a	.959a	.884a
	Average measures	0.968	0.979	0.938

Table 4: ICC result

5. RESULTS

Across the two tasks in L1 Vietnamese and L2 English, *attitude* constituted the major portion of the three *appraisal* categories. Student writers tended to express their emotions a lot when explaining the meaning of the flower narcissus, as well as arguing about how a certain lifestyle makes people happy. Both assertive claims and multi-voiced arguments were employed to convey the writer’s stance across the two tasks, occupying more than one-third of the overall evaluative resources. Around one-fifth of assessments heightened the actual attitudinal meaning by scaling or sharpening the writer’s instantiation (see Figure 6).

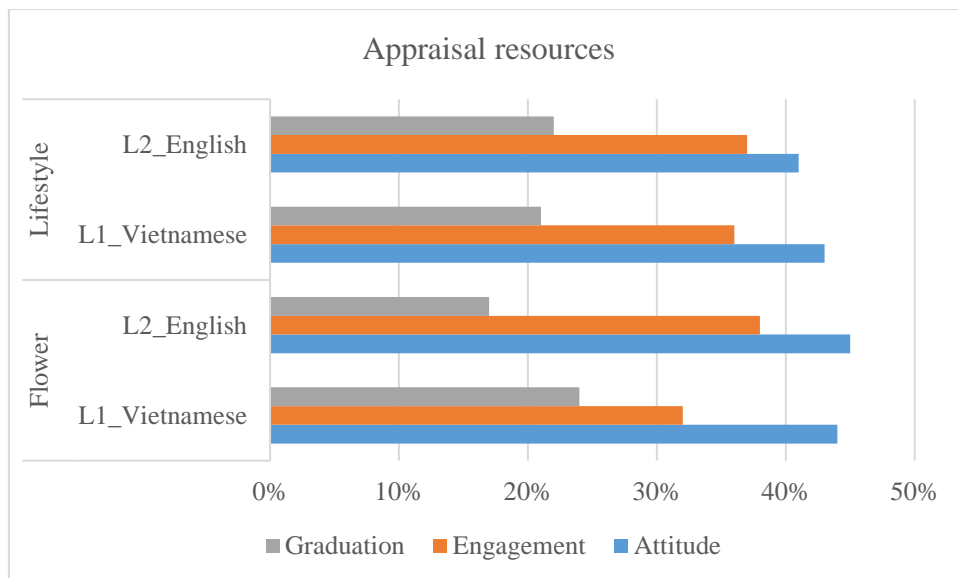


Figure 6: Distribution of *appraisal* resources

5.1. Attitude

For *attitude* resources in the first task, ‘flower’, shown in Table 4, *appreciation* accounted for the most annotations totalling around 50 per cent, while *affect* ranked second and *judgement* had the smallest proportion of annotations.

L1 Vietnamese and L2 English had a relatively balanced distribution of the three subcategories of *attitude*. Both L1 Vietnamese and L2 English writers preferred using *appreciation* to express their stance by aesthetically evaluating what is worth, what is made, and what is performed throughout their writing as in excerpts (27) and (28).

(27) Thủy tiên là một loài hoa **đẹp** [+*appreciation*], **thu hút** mọi ánh nhìn, loài hoa này thường **ngã thân hơi nghiêng xuống** [*neutral appreciation*] chứ **không thẳng** [*neutral appreciation*] như những loài hoa khác. // *The narcissus is a beautiful* [+*appreciation*] flower that *attracts* [+*appreciation*] all eyes, and this flower usually *leans its body down slightly* [*neutral appreciation*] but is *not as straight as* [*neutral appreciation*] other flowers. (L1VN)

(28) It's had such a **special** [+*appreciation*] beauty that everyone falls in love with it. [L2EN]

Student writers shared things in common in L1/L2 when talking about the ‘flower’. Evaluation mostly referred to how students emotionally reacted towards the flower or how they reflected on its physical appearance in general (excerpts (29) and (30)).

(29) Bản thân bạn dù **xấu** [+*appreciation*] **đẹp** [+*appreciation*] như thế nào thì đều **đáng trân trọng** [+*appreciation*] và hãnh diện cũng không nên quá đề cao bản thân mà nên quan tâm những người khác xung quanh bạn. // *You, no matter how ugly* [+*appreciation*] or *beautiful* [+*appreciation*] you become, are both *worthy* [+*appreciation*] and proud of yourself, [and] you should not overestimate yourself but should care about others around you. (L1VN)

(30) Daffodils are **beautiful** [+*appreciation*], **luxurious** [+*appreciation*], and **lovely** [+*appreciation*]. (L2EN)

While L1 Vietnamese writing mainly employed affection resources when discussing the flower, L2 English authors preferred a combination of affection, interest and pleasure (see excerpts (31) and (32)). Additionally, the emotions in both languages were primarily non-authorial although L2 writing made a few attempts to show their authorial evaluation as in excerpts (33) and (34) below. Concerning the valence and explicitness of attitudinal evaluation, there was no noticeable variation across L1 and L2 writing, in which positive values (excerpts (31) and (32) accounted for over 65 per cent and inscription occupied

over 93 per cent (excerpts (27)–(34)). Neutrality (excerpt (27)), invocation (excerpt (33)) and negativity (excerpt (34)) constituted under 20 per cent, under ten per cent and under four per cent respectively.

(31) Dù không màu sắc như hoa hồng, hoa cúc chỉ có thể là màu vàng và màu trắng nhưng nó rất **yêu** [+affect] vẻ đẹp riêng của mình như chàng Narziss **yêu** [+affect] vẻ đẹp của chàng và không để tâm đến ai cả. // *Although it is not as colorful as roses or chrysanthemums and can only be yellow and white, it loves [+affect] its own beauty very much as Narziss loves [+affect] the beauty of his and doesn't care about anyone.* (L1VN)

(32) When we have **admired** [+affect] the beauty of these flowers, we must be **passionate** [+affect] [...] Narziss is a great man who makes goddesses **love** [+affect] him, and daffodils make beautiful girls become **passionate** [+affect] about them and **love** [+affect] them. (L2EN)

(33) Nên “thủy” gắn liền với nước, và “tiên” do bông hoa xinh đẹp đến ngỡ ngàng và **ta không thể cưỡng lại sắc đẹp của nó** [t, +authorial]. // *Therefore, “thuy” is associated with water, and “tien” is because of the fact that the flower is surprisingly beautiful and we cannot resist its beauty [t, +authorial].* (L1VN)

(34) The writer wants to show that **we shouldn't be attracted to** [-authorial] beauty in order to not to ignore simple things. (L2EN)

Assessments of ethical or moral standing were frequently used by L1 Vietnamese and L2 English writers. Especially, student writers adopted this resource to mainly criticise the negativity of the flower or the beauty associated with the flower, such as arrogance, pride, selfishness or coldness (excerpts (35) and (36)). The other types of *judgement* were seldom used.

(35) Nhưng nó lại **quá tự cao** [-prop] vào nhan sắc của mình. // *However, it overvalues [-prop] its beauty.* (L1VN)

(36) He is **selfish** [-prop] and **cold** [-prop], so he cannot love everyone. [L2EN]

Regarding attitudinal variation, students' writing in Vietnamese seems more evaluative than that in English. Specifically, in Vietnamese paragraphs, writers tended to add more adjectives expressing their *attitude*, while in L2 English the flower and responsibilities were not described via adjectives (excerpts (37)–(40))

(37) Thủy tiên là một loài hoa **dại** [neutral appreciation] mọc ven hồ [...] // *The narcissus is a wild [neutral appreciation] kind of flowers that grows along the lake [...]* (L1VN)

(38) Thuy Tien is a flower growing near the lake. (L2EN)

(39) [...] không phải lo lắng về trách nhiệm **bất đắc dĩ** [-affect]. // [...] *without worrying about **reluctant** [-affect] responsibilities.* (L1VN)

(40) [...] without responsibility. (L2EN)

Additionally, writers used different words between English and Vietnamese revealing various *attitudes*. Thus, for example, under the same subcategory of *attitude* (*appreciation* and *affect*), different specific *appraisal* features were used to express the beauty of the flower or its feeling (excerpts (41) to (44)). In L2 English, writers preferred to appreciate the flower while in L1 Vietnamese they expressed how the flower felt, as in excerpts (45) and (46).

(41) [...] nó có một vẻ đẹp rất **quyến rũ** [+appreciation: reaction-impact] mà bất cứ ai cũng say mê đắm đuối. [...] *it has a very **seductive** [+appreciation: reaction-impact] beauty that anyone can be passionately infatuated with.* // (L1VN)

(42) It's had such a **special** [+appreciation: social valuation] beauty that everyone falls in love with it. [L2EN]

(43) [...] đặc biệt bông hoa luôn hướng xuống dưới giống như đang rất **buồn bã** [-affect: unhappiness-misery]. // [...] *especially the flower is always facing down as if it is very **sad** [-affect: unhappiness-misery].* (L1VN)

(44) Especially, the flowers are always looking down as [if] they are very upset [-affect: insecurity-disquiet]. (L2EN)

(45) Hoa thủy tiên là một loài hoa rất đặc biệt và là một loài hoa **thích** [+affect: happiness-affection] sự cô độc. // *The narcissus is a kind of flowers that is very special and is a kind of flowers that **likes** [+affect: happiness-affection] its loneliness.* (L1VN)

(46) Thuy Tien is a special flower, [and] it's also a **lovely** [+appreciation] flower. (L2EN)

Task	Flower				Lifestyle			
	L1_Vietnamese		L2_English		L1_Vietnamese		L2_English	
Feature	N	Percent	N	Percent	N	Percent	N	Percent
Attitude-type	N=77		N=59		N=111		N=105	
affect	28	36%	24	41%	47	42%	43	41%
judgement	13	17%	8	13%	29	26%	34	32%
appreciation	36	47%	27	46%	35	32%	28	27%
Non/authorial evaluation	N=28		N=24		N=47		N=43	
authorial	1	4%	4	17%	14	30%	15	35%
non-authorial	27	96%	20	83%	33	70%	28	65%
Valence	N=77		N=59		N=111		N=105	
positive-attitude	51	66%	41	69%	72	65%	79	75%
negative-attitude	17	22%	11	19%	35	31%	24	23%
neutral/ambiguous	9	12%	7	12%	4	4%	2	2%
Explicitness	N=77		N=59		N=111		N=105	
inscribed	72	94%	57	97%	108	97%	103	98%
invoked	5	6%	2	3%	3	3%	2	2%

Table 5: Distribution of *attitude*

For the ‘lifestyle’ topic (Table 5), the most frequently employed *appraisal* resource across L1 and L2 was *affect*. However, L1 Vietnamese texts used more *appreciation* than *judgement* while L2 English texts preferred evaluating behaviour to appreciating things. Concerning *appreciation*, L1 Vietnamese writing had the tendency to evaluate what was socially valued when arguing about the importance of beliefs and purposes in life. This evaluative tool outnumbered their affective responses to or assessments of the composition of something. On the contrary, L2 English writers reacted to and produced their own social valuation in relatively equal attempts (excerpts (47) and (48)).

(47) Cuộc đời của mỗi con người chính là một chuỗi liên tiếp những **khó khăn** và **thử thách** [-*appreciation*], do đó **quan trọng** [+*appreciation*] là bạn lựa chọn vượt qua nó như thế nào. // *The life of every human being is a series of difficulties and challenges* [-*appreciation*], so the **important** [+*appreciation*] thing is how you choose to overcome it. (L1VN)

(48) The second man lives because of waiting for a **beautiful** [+*appreciation*], **good** [+*appreciation*] life in the future. (L2EN)

With respect to *affect*, writers seemed to utilise the same strategies across L1/L2 production in conveying their emotions. Talking about what makes life meaningful, they often employed positive evaluative resources such as cheering, trusting, confidence, and relaxation (excerpts (49) and (50)). Negative assessments (excerpts (51) and (52)) tended to mention about the obstacles and dissatisfaction in people's life, accounting for slightly over one-third of the overall attitudinal resources in L1 Vietnamese texts while L2 English writing adopted fewer negative emotions. Authorial stance (excerpts (49)–(51)) across the two languages in this task seemed to increase in comparison with the first task, while explicit evaluation was still dominant.

(49) Từ đó bản thân ta sẽ **vui vẻ** [+affect; +authorial], **lạc quan** [+affect; +authorial] và ngày càng yêu đời hơn. // *Since then, we will be **happy** [+affect; +authorial], **optimistic** [+affect; +authorial] and love life more and more.* (L1VN)

(50) Therefore, living **happily** [+affect; +authorial] and making ourselves feel **happy** [+affect; +authorial] is enough. (L2EN)

(51) Vậy tại sao chúng ta không thể tập cho bản thân chúng ta có cái nhìn tích cực hơn mà lại **than thở** [-affect, +authorial] **trách móc** [-affect, +authorial] tại sao chuyện như thế này, như thế kia. // *Therefore, why can't we train ourselves to have a more positive outlook? But we just **complain** [-affect, +authorial] and **blame** [-affect, +authorial] why things happen like this or like that.* (L1VN)

(52) That is the reason why people feel **unhappy** [-affect], and they are **stressful** [-affect] in their life. (L2EN)

Relating to the overall pattern of evaluation of behaviour, writers from both groups made the most use of judgments of the *capacity*, *normality*, and *propriety* of the subject matter, although L2 English writers did so more frequently (excerpts (53) and (54)). Assessments of how truthful or honest a person is were not employed in the task 'lifestyle', presumably because of a lack of common ground required for such assumptions.

(53) Lý tưởng là những nguyên tắc do ta **đặt ra** [+judgement] và **cố gắng thực hiện** [+judgement] trong cuộc sống. // *Ideals are the principles we **set out** [+judgement] and **try to implement** [+judgement] in life.* (L1VN)

(54) Living with the ideal **helps** [+judgement] us overcome the fear, unexpectation, and **have more responsibilities** [+judgement] in life. (L2EN)

The vocabulary and structures used between English and Vietnamese versions in the second task apparently indicate that students used attitudinally identical terms (excerpts (55)–(58)).

(55) Sau khi đọc đoạn trích, tôi nhận ra vai trò **quan trọng** [+appreciation] của niềm tin trong cuộc sống. // *After reading the excerpt, I realize the **important** [+appreciation] role belief plays in life.* (L1VN)

(56) After reading the paragraph below, I notice that belief plays **important** [+appreciation] role in life. (L2EN)

(57) Vì thế, dù bạn **đang gặp khó khăn** [neutral judgement] hay có những suy nghĩ **tiêu cực** [-appreciation] thì xin hãy tin rằng ngày mai sẽ **tốt hơn** [+appreciation] hôm nay. // *Hence, although you are **struggling** [neutral judgement] or have **negative** [-appreciation] thoughts, please believe that tomorrow will be **better** [+appreciation than today.* (L1VN)

(58) Therefore, when you meet **difficulties** [-appreciation] or have **bad** [-appreciation] thoughts, please believe that tomorrow is **better** [+appreciation and force: upscale] than today. (L2EN)

5.2. Engagement

Regarding the first task ('flower'), as illustrated in Table 6, L1 Vietnamese texts contained more heteroglossic resources than monoglossic ones. However, L2 English had slightly more bare assertions than multi-voiced arguments.

Task	Flower				Lifestyle			
Feature	L1_Vietnamese		L2_English		L1_Vietnamese		L2_English	
	N	Percent	N	Percent	N	Percent	N	Percent
Engagement-type	N=56		N=50		N=93		N=97	
mono-glossic	18	32%	27	54%	20	22%	15	15%
hetero-glossic	38	68%	23	46%	73	78%	82	85%
Heteroglossic-type	N=38		N=23		N=73		N=82	
contract	25	66%	14	61%	39	53%	37	45%
expand	13	34%	9	39%	34	47%	45	55%
Contract-type	N=25		N=14		N=39		N=37	
disclaim	17	68%	7	50%	16	41%	15	41%
proclaim	8	32%	7	50%	23	59%	22	59%
Disclaim-type	N=17		N=7		N=16		N=15	
deny	8	47%	5	71%	9	56%	10	67%
counter	9	53%	2	29%	7	44%	5	33%
Proclaim-type	N=8		N=7		N=23		N=22	
concur	1	13%	1	14%	2	9%	1	5%
pronounce	2	25%	1	14%	3	13%	2	9%
endorse	0	0%	0	0%	0	0%	0	0%
justify	5	63%	5	71%	18	78%	19	86%
Expand-type	N=13		N=9		N=34		N=45	
entertain	11	85%	8	89%	30	88%	40	89%
attribute	2	15%	1	11%	4	12%	5	11%

Table 6: Distribution of *engagement*

Under the umbrella of *heteroglossia*, contraction was preferred to expansion across L1 and L2 texts. Specifically, while writers using L2 English applied as much *disclaim* as *proclaim*, in L1 Vietnamese writing disclamation outweighed proclamation. Within the disclamation subcategory, writers using L1 Vietnamese had a relatively balanced usage of denial and countering. When writing in L2 English, writers relied mainly on denying. Regarding the distribution of *proclaim*, both L1 Vietnamese and L2 English texts did not adopt endorsement, while of the remaining three *proclaim* types, writers in both languages employed justification more frequently than the other types (excerpts (59) and (60)).

- (59) Cũng chính vì [*justify*] chàng tự say mê sắc đẹp của mình một cách thái quá nên dẫn đến cái chết thương tâm. // It is also **because** [*justify*] he is too infatuated with his own beauty that leads to tragic death. (L1VN)

- (60) **For this reason** [*justify*], he felt confident in his beauty and dealt with his death. (L2EN)

Regarding expansive resources, modality of usuality and probability via the modal verb *can* contributed to one-fourth of the total entertainment instances across the L1 and L2 texts (e.g. excerpt (61)). Other resources included personalisation, modalised cause *if* and cases of obligatory modality. A few external sources as attribution were also employed to open alternatives for other viewpoints (excerpt (62)).

- (61) From the story, people **can** [*entertain*] know the source of “Narcissus.” (L2EN)

- (62) Thủy tiên, **theo cách gọi tên hoa** [*attribute*] có nghĩa là “tiên nước”, vị tiên nơi thủy cung. // *Thuy Tien, according to the way naming flowers* [*attribute*], has the meaning of “water fairy”, the fairy in the underwater imperial palace. (L1VN)

There are instances where L2 English paragraphs included more features of *engagement*, while L1 Vietnamese equivalents included fewer features by count, but the features that were employed spanned multiple words (excerpt (63)). Furthermore, there were instances in which *engagement* resources were present in L2 English but not in L1 Vietnamese (excerpts (64) and (65)). Besides, the authorial stance in excerpt (64) was strongly emphasised while in excerpts (65)–(67) non-authorial treatment was employed.

- (63) **From the story** [*attribute*], Narziss is a pretty boy. (L2EN)

- (64) When we have admired the beauty of these flowers, we **must** [*entertain*] be passionate. (L2EN)

- (65) Khi chiêm ngưỡng vẻ đẹp của hoa, người ta càng thêm say đắm. // *When admiring the beauty of the flower, people are more and more infatuated.* (L1VN)

- (66) The beauty can make **everything** infatuated with it. (L1VN)

- (67) From the story of Narziss, we can see that **narcissus** is very beautiful. (L2EN)

In relation to the second task, ‘lifestyle’, L2 English texts employed the most heteroglossic resources, though the average word counts of these paragraphs appeared to be smaller than their L1 Vietnamese counterparts. The distribution of specific

heteroglossic types between L1 Vietnamese and L2 English was similar. Both lacked the usage of endorsement, as in the first task.

While writers were still in favour of contracting options for alternative voices when writing in their L1, writers using L2 preferred expanding chances for multi-voiced argumentation. However, both L1/L2 texts shared several things in common in highlighting the use of proclaiming, denying, justifying, and entertaining. Particularly, the authorial voice revealed in L1 texts seemed adding emphasis to the overall stance while the pronouncement in the L2 was adopted to doubtlessly reinforce the cohesion of the writing (excerpt (68) and (69)). Being dialogically open, both L1/L2 texts favoured personalisation in their claiming (excerpt (70) and (71)), leaving a scattered range of other resources related to probability, usuality, obligation, and modalised cause. Similarly, to task one, attribution was also employed around ten per cent across the two languages.

(68) **Trong thực tế** [*pronounce*], sự lạc quan hay sự thoải mái về mặt tinh thần là một vũ khí giúp ta có thể vượt qua những trở ngại trong cuộc sống cũng như có thể là một vị thuốc tốt nhất để chữa trị những căn bệnh hiểm nghèo. // **In fact** [*pronounce*], *optimism or mental comfort is a weapon that can help us overcome the obstacles in life and can be the best medicine to cure serious illnesses.* (L1VN)

(69) **Additionally** [*pronounce*], there are many things you need to think and scare. (L2EN)

(70) Sau khi đọc đoạn trích, **tôi nhận ra** [*entertain*] vai trò quan trọng của niềm tin trong cuộc sống. // *After reading the excerpt, I realize* [*entertain*] *the important role belief plays in life.* (L1VN)

(71) **I think** [*entertain*] we shouldn't worry about anything, though our lives still have so many difficulties and sadness. (L2EN)

Moreover, L2 English writing's employment of *proclaim* and *entertain* (excerpt (72)) was different from the L1 Vietnamese writing's adoption of only *proclaim* (excerpt (73)). Also, L2 English tendency to favour medium modality of obligation *should* contradicted L1 Vietnamese preference of its high obligatory degree *must* (excerpts (74) and (75)). One interesting fact is that L2 English writing used redundancy while there was no equivalent in L1 Vietnamese as in excerpts (76) and (77). This goes against claims that Vietnamese students of English seem to transfer their L1 redundancy to their L2 English such as in *Although ..., but ...* (Ho 2011: 183).

(72) **Because** [*proclaim: justify*] **I believe** [*entertain*] in “After raining, the sun is rising”, I come over my challenge. (L2EN)

(73) Cũng **vì** [*proclaim: justify*] có niềm tin “sau cơn mưa trời lại sáng” thì có bao nhiêu khó khăn có đáng là gì. // Also **because** [*proclaim: justify*] there is the belief “after the rain, it will be sunny again”, it’s worth facing difficulties. (L1VN)

(74) Bản thân của mỗi người **phải** [*entertain*] biết tìm ra cho mình một lý tưởng sống. // Each person **must** know how to find out for himself or herself an ideal of life. (L1VN)

(75) We ourselves **should** [*entertain*] find an ideal. (L2EN)

(76) Sau khi đọc đoạn trích trên, **theo em nghĩ** [*entertain*], sống vui vẻ và hạnh phúc là sống không lo lắng sợ hãi, không chờ đợi [...] // After reading the above excerpt, **in my opinion** [*entertain*], living cheerfully and happily is living without worry or fear, without waiting [...] (L1VN)

(77) After I read the passage, **in my opinion I think** [*entertain*] to live for a happy and good life is to live without worry, without scare, without waiting for something [...] (L2EN)

Matching L1/L2 heteroglossic resources were noted in the use of entertainment and justification in excerpts (55) to (58) in Section 5.1. For example, **tôi nhận ra** [*entertain*] and **I notice** [*entertain*] as well as **vì thế** [*justify*] and **therefore** [*justify*] were perfectly matched as if they were translated from L1 into L2. Another similarity identified in the second task, ‘lifestyle’, is that the sequence *I think* in English and its counterpart in Vietnamese, *Tôi nghĩ*, appeared five times in both ILVs.

5.3. Graduation

Force was dominantly used in comparison with *focus* in both tasks. As shown in Table 7, upscale gradability was in major usage and sharpening was preferable. However, there existed some variation across L1 Vietnamese and L2 English texts.

Task	Flower				Lifestyle			
	L1_Vietnamese		L2_English		L1_Vietnamese		L2_English	
	N	Percent	N	Percent	N	Percent	N	Percent
Graduation-type	N=41		N=22		N=53		N=57	
force	34	83%	19	86%	44	83%	46	81%
focus	7	17%	3	14%	9	17%	11	19%
Scale	N=34		N=19		N=44		N=46	
upscale	30	88%	16	84%	42	95%	46	100%
downscale	4	12%	3	16%	2	5%	0	0%
Focus-type	N=7		N=3		N=9		N=11	
soften	0	0%	0	0%	1	11%	0	0%
sharpen	7	100%	3	100%	8	89%	11	100%

Table 7: Distribution of *graduation*

Regarding the task ‘flower’, the majority of *force* focused on intensification with some exceptions of quantification. L1 writing adopted only one instance of quantifying by using time distribution (excerpt (78)) while L2 texts employed numbers (excerpt (79)). Additionally, both degree or quality and vigour or process were intensified in relatively balanced distribution although in L1 vigour was slightly dominant (excerpt (66), repeated here as (80)) and in L2 degree was favoured (excerpt (67), repeated here as (81)).

(78) Không thể phủ nhận rằng vẻ đẹp của thủy tiên làm người ta muốn sở hữu **mãi** [*force: upscale*]. // *There is no denying that the daffodil's beauty causes ones to desire to have it forever* [*force: upscale*]. (L1VN)

(79) And “tien” has an amazing beauty, so **all** [*force: upscale*] people are attracted to it. (L2EN)

(80) The beauty can make everything **infatuated** [*force: upscale*] with it. // (L1VN)

(81) From the story of Narziss, we can see that narcissus is **very** [*force: upscale*] beautiful. (L2EN)

Specifically, no comparatives and superlatives were used to intensify what was to be conveyed in the first task. Instead, intensifiers such as *very/rất* and *too/quá* were over-employed in both L2 English and L1 Vietnamese paragraphs (excerpts (82) and (83)). Especially, no equivalent between L1 Vietnamese and L2 English can be noted in excerpts (84) and (85). The English version used resources of *graduation* to emphasise the act of lonely but selfish living whereas the Vietnamese one preferred using emotions to praise the relaxation of living.

(82) Tác giả cũng muốn cho ta thấy đừng **quá** [*upscale*] chìm đắm vào cái đẹp để rồi quên đi những thứ giản dị bình thường. // *The author also wants to show us not to be **too** [*upscale*] immersed in beauty and then forget about the ordinary simple things.* (L1VN)

(83) But it's **too** [*upscale*] proud of its beauty. (L2EN)

(84) Hoa thủy tiên không quan tâm những gì xung quanh nó, cứ **bình thản** [*+affect: security-quiet*] sống. // *The narcissus does not care about what is around it but lives **at ease** [*+affect: security-quiet*].* (L1VN)

(85) Thuy Tien flowers don't care anything around them, [and] they just [*upscale*] live.

Regarding the second task, 'lifestyle', quantification was employed more frequently in L2 via the use of number (excerpt (86)), and vigour was intensified three times as much as degree in L1. Only one instance of softening value was recorded in L1 writing (excerpt (87)). In particular, L2 English demonstrated the largest adoption of upscale grading. However, L1 Vietnamese allowed for use of *focus* and downscale attitudinal values (excerpt (87)). The over-representation of *force* as well as the under-adoption of *focus* in L2 English writing might result from lack of linguistic devices.

(86) It helps us to develop and discover **many** [*force: upscale*] new things in our live. (L2EN)

(87) **Có lẽ** [*force: downscale*] **một phần** [*focus: soften*] do bệnh nhân ấy đến từ một vùng quê nên không hiểu về những gì bác sĩ nói về căn bệnh của mình nên có suy nghĩ mình sẽ khỏi bệnh. // ***Perhaps** [*force: downscale*] **partly** [*focus: soften*] because the patient was from the countryside, he did not understand what the doctor said about his illness, and he thought he would be cured.* (L1VN)

Unlike the first task, comparatives and superlatives to indicate isolating intensification were over-represented in both L1 Vietnamese and L2 English in the second task (excerpts (88) and (89)).

(88) Từ đó bản thân ta sẽ vui vẻ, lạc quan và **ngày càng** yêu đời **hơn** [*upscale*]. // *Since then, we will be happy, optimistic and love life **more and more** [*upscale*].* (L1VN)

(89) It's the **best** [*upscale*] medicine to treat many diseases. (L2EN)

6. DISCUSSION

The present study has explored the three primary domains of *appraisal* to discover how evaluative resources were employed in native Vietnamese and intermediate L2 English paragraphs written by the same writers, across the same tasks and under the same conditions. Based on CIA² model and using a corpus-informed approach, these two ILVs were compared and contrasted with findings intended to help researchers and English language teachers identify areas where their students need improvements.

The overall pattern of *appraisal* resources in both tasks across L1 and L2 is characterised by the use of *attitude*, *engagement*, and *graduation* resources in descending order of frequency. Writers in both languages had a particular tendency to express their emotions, assess phenomena and judge behaviour in generalising the flower's beauty and lifestyle in connection with people's well-being. Multi-voiced reasoning was preferred in general, with the exception of L2 English texts in the first task. These particular writers frequently adopted bare assertions when discussing about the flower. On the whole, intensifiers were mostly employed to emphasise the intended meaning. This general distribution of *appraisal* resources was in line with Chung (2018) in that L2 English writers from Vietnamese background in Chung (2018) expressed a lot of their personal feelings when reasoning. However, the findings of the present study were different from Lam and Crosthwaite (2018), in that the native English writers and L2 English writers from an L1 Cantonese background in that study tended to make frequent use of personal claims and alternative voices in their persuasive argumentation. As Lam and Crosthwaite dealt with Cantonese, further investigation is required to examine whether speakers of Vietnamese adopt this style in L2 English based on intrusion from their L1 or a lack of required L2 resources.

Within the *attitude* domain, L1 Vietnamese used such resources more frequently than in L2 English, mainly as the writers' L1 Vietnamese production tended to add adjectives to modify nouns (excerpts (37)–(40)). This echoed Phan's (2011) analysis, which indicated that Vietnamese writing usually adopted 'flowery style'. Regarding the first task, the present research discovered the prevailing use of *appreciation*, which was in harmony with expert writers' L1 English production in Hood (2004) and in agreement with L1/L2 English texts in Lam and Crosthwaite (2018). However, for the second task, *affect* was predominantly used. This finding is in opposition to Derewianka (2007), who indicated that low proficiency writers had the tendency to overuse emotions and

assessment of behaviour in their evaluation. However, in our study the same writers performed both tasks, and so the variation in attitudinal evaluation was less likely to have been caused by proficiency, as was the case with Derewianka's study. Hood (2004) also showed L2 English writers are often inclined to embed personalised treatment through emotions, yet this was not found in the first task. Our findings, based on having the same writers perform both tasks, may suggest that the topic may be more responsible for this difference in distribution.

As for the *engagement* category, L2 writing in the first task favoured *contractive* resources in stancetaking over those seen in L1 texts, echoing Coffin and Hewings (2004) and Lam and Crosthwaite (2018). One explanation may be that L2 learners possess a limited range of vocabulary and structures for expansive resources related to the specific topic. L2 writing in the second task, however, was more open for alternative viewpoints, which was in line with Ho (2011) who found speakers of Vietnamese frequently adopted expansive resources in their arguments. Again, the nature of the task may be ultimately responsible for this variation, over issues related to L1/L2 differences. The major presence of justification in the first L2 task is found in line with Lam and Crosthwaite (2018) where L2 English writers from L1 Cantonese (Hong Kong) backgrounds provided reasons for their propositions through overuse of transitions and frame markers (*because, since, the reason for, etc.*). Through justifying their claims, L2 writers of Vietnamese background tried to show their reasoning skill in the target language. That heteroglossic employment was preferable in L2 texts in the second task, 'lifestyle' is consistent with the findings of Ho (2011), Lancaster (2011), and Lam and Crosthwaite (2018) in that L2 students prefer to engage in explicit dialogue in their writing when conveying their ideas while opening space for discussion with readers, at the expense of monoglossic asides or reflection.

In terms of *graduation*, both L1/L2 tasks reflected the dominance of *force* although more *focus* was placed on use in L1 than in L2, as already shown by Lam and Crosthwaite (2018). *Upscale* resources were over employed to heighten the intended meaning while *focus* was under adopted. Particularly, in the second task, upscale evaluation made the largest contribution in L2 texts, which is similarly found in L1 English writing in Lam and Crosthwaite (2018). Overuse of *force* in L2 English and lack of sharpening or softening non-gradable attitudinal meanings may be beyond their L2 language proficiency, as also indicated in Lam and Crosthwaite (2018). The unbalanced *force-focus*

adoption in L2 texts suggested that Hood (2006) and Derewianka (2007) are right in recommending both the equipping of linguistic resources and managing these resources to help L2 learners better deploy the scaling and sharpening/softening of their attitudinal evaluation.

Our hypothesis that our students tended to directly translate their L1 Vietnamese use of *appraisal* into L2 English writing is partially supported by the data. Identical evaluative resources in excerpts (55)–(58) were presented when writers appreciated the importance of belief in people's lives and assessed their obstacles. L1 is found to be useful in almost all writing stages from planning, reformulating to revising and monitoring (Sasaki 2002), while nearly half of skilled L2 writers in Beare and Bourdages (2007) adopted translation from L1 as their primary writing strategy. Lam and Crosthwaite (2018) also discovered cases of L1 transfer when L2 writers tried to make extensively positive or negative evaluation. However, the matching of *appraisal* resources in those excerpts did not hinder the conveyed meaning, suggesting learners were able to positively transfer *appraisal* resources from L1 to L2 writing, and L2 teachers should be cognisant of the funds of knowledge that L2 writers can bring from their L1.

7. CONCLUSION

In summary, the present study has presented a range of similarities and differences in *appraisal* resource usage between L1 Vietnamese and L2 English. English language teachers should pay more attention to which *appraisal* resources their Vietnamese students tend to overuse or underuse in their English writing to help them better adjust future instruction and materials preparation. In particular, successful argumentative writing has to relatively balance *expansive* and *contractive* alternatives in which the former should outweigh the latter (Lancaster 2011). In line with Hood (2004) and Lam and Crosthwaite (2018), this research suggests L2 English writers need to be provided with these interpersonal resources in improving their writing. However, as L1 Vietnamese writing is also filled with *appraisal*, L2 writing instructors are advised to maximise positive transfer of L1 features in L2 texts where appropriate. The value of the present study is that corpus analyses of this nature help to reveal which L1 features positively transfer in L2 production and which do not do so as readily.

In terms of limitations, the small scale of this research makes it difficult to generalise findings beyond the sample surveyed. Also, a one-week time gap between the two writing tasks might indicate the possibility that student writers could transfer what they wrote between L1 and L2, although this is unlikely. Further research involving more diatypic and dialogic variables needs to be considered for full use of the CIA² approach (Granger 2015). Moreover, as this is a descriptive discourse analysis paper covering a small dataset, we have not sought to compare writers' production of *appraisal* resources across the two writing tasks through inferential statistical, providing instead a descriptive overview of the writers' production across both tasks in the results section. We are working on a follow-up study that will explore task effects in more quantitative detail. Finally, our aim in this paper is to provide an overview of *appraisal* using the whole framework, which is possible given the relatively small dataset. A more detailed analysis of L1/L2 writer's production across individual *appraisal* categories is the subject of forthcoming research.

REFERENCES

- Bailey, Stephen. 2006. *Academic Writing: A Handbook for International Students*. Oxfordshire: Routledge.
- Beare, Sophie and Johanne Bourdages. 2007. Skilled writers' generating strategies in L1 and L2: An exploratory study. In Mark Torrance, Luuk van Waes and David Galbraith eds. *Writing and Cognition: Research and Applications*. Amsterdam: Elsevier, 151–161.
- Biber, Douglas and Edward Finegan. 1989. Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text* 9/1: 93–124.
- Bley-Vroman, Robert. 1983. The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning* 33: 1–17.
- Chafe, Wallace L. and Johanna Nichols eds. 1986. *Evidentiality: The Linguistic Coding of Epistemology*. Norwood, NJ: Ablex.
- Chung, Thuy T. 2018. Discovering interpersonal stance in EFL Students' writing. Paper presented at the 14th Annual Cam TESOL Conference on English Language Teaching, Phnom Penh, Cambodia, 10–11 February 2018.
- Coffin, Caroline and Ann Hewings. 2004. IELTS as preparation for tertiary writing: Distinctive interpersonal and textual strategies. In Louise J. Ravelli and Robert A. Ellis eds. *Analysing Academic Writing: Contextualised Frameworks*. London: Continuum, 153–171.
- Derewianka, Beverly. 2007. Using *appraisal* theory to track interpersonal development in adolescent academic writing. In Anne McCabe, Mick O'Donnell and Rachel Whittaker eds., 142–165.
- Geng, Yifan and Sue Wharton. 2016. Evaluative language in discussion sections of doctoral theses: Similarities and differences between L1 Chinese and L1 English writers. *Journal of English for Academic Purposes* 22: 80–91.
- Goffman, Erving. 1981. *Forms of Talk*. Oxford: Oxford University Press.

- Granger, Sylviane. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In Karin Aijmer, Bengt Altenberg and Mats Johansson eds. *Languages in Contrast. Papers from a Symposium on Text-Based Cross-Linguistic Studies*. Lund: Lund University Press, 37–51.
- Granger, Sylviane. 2015. Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1/1: 7–24.
- Ho, Vu L. 2011. *Non-native Argumentative Writing by Vietnamese Learners of English: A Contrastive Study*. Washington, DC: The Georgetown University dissertation.
- Hood, Susan. 2004. Managing attitude in undergraduate academic writing: A focus on the introductions to research reports. In Louise J. Ravelli and Robert A. Ellis eds. *Analysing Academic Writing: Contextualized Frameworks*. London: Continuum, 24–44.
- Hood, Susan. 2006. The persuasive power of prosodies: Radiating values in academic writing. *Journal of English for Academic Purposes* 5/1: 37–49.
- Hunston, Susan and Geoff Thompson. 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.
- Hyland, Ken. 2005. Stance and engagement: A model of interaction in academic discourse. *Discourse Studies* 7/2: 173–192.
- Koo, Terry. K. and Mae Y. Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* 15/2: 155–163.
- Labov, William. 1984. Intensity. In Deborah Schiffrin ed. *Meaning, Form, and Use in Context: Linguistic Applications*. Washington: University of Georgetown Press, 43–70.
- Lam, Suet L. and Peter Crosthwaite. 2018. *Appraisal* resources in L1 and L2 argumentative essays: A contrastive learner corpus-informed study of evaluative stance. *Journal of Corpora and Discourse Studies* 1/1: 8–35.
- Lancaster, Zac. 2011. Interpersonal stance in L1 and L2 students' argumentative writing in economics: Implications for faculty development in WAC/WID programs. *Across the Disciplines* 8/4: 1–23.
- Martin, James R. and Peter R. R. White. 2005. *The Language of Evaluation: Appraisal in English*. New York: Palgrave Macmillan.
- McCabe, Anne, Mick O'Donnell and Rachel Whittaker eds. 2007. *Advances in Language and Education*. London: Continuum.
- McEnery, Tony and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
- Ngo, Thu and Len Unsworth. 2015. Reworking the *appraisal* framework in ESL research: Refining attitude resources. *Functional Linguistics* 2/1: 1–24.
- Ochs, Elinor. 1989. Introduction. *Text – Interdisciplinary Journal for the Study of Discourse* 9/1: 1–5.
- O'Donnell, Mick. 2016. *UAM CorpusTool* 3.3. <http://corpustool.com/download.html> (8 August, 2018.)
- Phan, Ha L. 2011. The writing and culture nexus: Writers' comparisons of Vietnamese and English academic writing. In Ha L. Phan and Bradley Baurain eds. *Voices, Identities, Negotiations and Conflicts: Writing Academic English across Cultures*. Bingley: Emerald, 23–40.
- Sasaki, Miyuki. 2002. Building an empirically-based model of EFL learners' writing processes. In Gert Rijlaarsdam, Sarah Ransdell and Marie-Laure Barbier eds. *New Directions for Research in L2 Writing*. Dordrecht: Kluwer, 49–78.

- Simpson, Paul. 1993. *Language, Ideology and Point of View*. London: Routledge.
- Swain, Elizabeth. 2007. Constructing an effective ‘voice’ in academic discussion writing: An *appraisal* theory perspective. In Anne McCabe, Mick O’Donnell and Rachel Whittaker eds., 166–184.
- Tran, Van T. H. 2011. *A Linguistic Study on Social Attitudes toward the Quality Issues of Postgraduate Education in Vietnam*. Wollongong, NSW: The University of Wollongong dissertation.
- Vo, Duc D. 2011. *Styles, Structure and Ideology in English and Vietnamese Business Hard News Reporting – A Comparative Study*. Adelaide, SA: The University of Adelaide dissertation.
- Vo, Trang N. T. 2017. Linguistic expression of judgment and appreciation in English and Vietnamese newspaper articles of social issues. Paper presented at the *3rd National Conference on Interdisciplinary Research in Linguistics and Language Education*, Hue, Vietnam, 22 November 2017.

Corresponding author

Peter Crosthwaite
 Room 510, School of Languages and Cultures
 Gordon Greenwood Building
 University of Queensland
 St. Lucia, Australia, 4072
 e-mail: p.cros@uq.edu.au

received: October 2020
 accepted: January 2021
 published online: February 2021

The *Process Corpus of English in Education*: Going beyond the written text

Gaëtanelle Gilquin
Université catholique de Louvain / Belgium

Abstract – The *Process Corpus of English in Education* (PROCEED) is a learner corpus of English which, in addition to written texts, consists of data that make the writing process visible in the form of keystroke log files and screencast videos. It comes with rich metadata about each learner, among which indices of exposure to the target language and cognitive measures such as working memory or fluid intelligence. It also includes an L1 component which is made up of similar data produced by the learners in their mother tongue. PROCEED opens new perspectives in the study of learner writing, by going beyond the written product. It makes it possible to investigate aspects such as writing fluency, use of online resources, cognitive phenomena like automaticity and avoidance, or theoretical modelling of the writing process. It also has applications for teaching, e.g. by showing students screencast video clips from the corpus illustrating effective writing strategies, as well as for testing, e.g. by establishing a corpus-derived standard of writing fluency for learners at a certain proficiency level.

Keywords – Learner corpus research; process learner corpus; writing process; keylogging; screencasting; metadata

1. INTRODUCTION: FROM WRITTEN PRODUCT TO WRITING PROCESS

The first electronic corpus ever, the *Brown Corpus*, was a corpus of written English. Since then, many corpora have been collected that represent written language. Among learner corpora, i.e. corpora consisting of language produced by foreign or second language (L2) learners, 64 per cent are made up of written texts only (and 12% of both written texts and spoken transcripts) according to the current version of the *Learner Corpora around the World* list maintained by the *Centre for English Corpus Linguistics* (2020). Examples of written learner corpora include the *International Corpus of Learner English*, the *Longman Learners' Corpus*, the *International Corpus of Crosslinguistic Interlanguage* or the *Written Corpus of Learner English*. These and other written corpora have yielded invaluable insights into writing: its lexico-

grammatical features, the way sentences and paragraphs are organised, how genres can be characterised linguistically, what errors writers tend to make, etc.

What these corpora give access to is the written product, that is, the final output of the writing act. Most written texts, however, go through several stages of editing and revision before they reach the final stage, when the text is offered to the reader. These intermediate states of the text are lost in a typical written corpus. The aim of the resource that is introduced in this article, the *PROcess Corpus of English in EDucation* (PROCEED),¹ is to make the whole writing process visible. To illustrate the difference between written product and writing process, one can consider example (1), a sentence taken from PROCEED and produced by a French-speaking learner of English. This sentence is the result of as many as twenty-eight different stages, as visible in PROCEED and as represented in (2), where strikethrough indicates text that has been deleted and the grey font shows a word in which one or several letters have been inserted.

- (1) Our actual society is dominted by technology and science. A lot of experiments concentrate lately on the effects of those new developments on the human being.

- (2) a. In
 b. ~~In~~
 c. Our moder
 d. ~~Our moder~~
 e. It is o
 f. ~~o~~
 g. nowadays a fact: our modern society is dominated by
 h. ~~It is nowadays a fact: our modern society is dominated by~~
 i. Our actual society is dominted by num
 j. ~~num~~
 k. techonology and scin
 l. ~~n~~
 m. encee
 n. ~~e~~
 o. . A lot of experiments concetrate lately to the effect of those
 p. ~~concentrate~~
 q. ~~to~~
 r. on
 s. effects
 t. ~~of those~~
 u. that it could cause for the human beings.
 v. ~~that it could cause for the human beings.~~

¹ See <https://uclouvain.be/en/research-institutes/ilc/cecl/proceed.html> (8 March, 2021.)

- w. that
- x. ~~that~~
- y. of those new devem
- z. ~~m~~
- aa. lopment on the human being.
- ab. developments

The intermediate stages reveal, among others, errors that have been corrected, for example *concentrate to* (2q) which has been replaced by *concentrate on* (2r), but also, more surprisingly, correct phrases that have been replaced by incorrect ones, as appears from the transformation of *modern society* into *actual society* (2h-i), with the use of a false friend (in French *actuel* means ‘current’). Although (2g) is not kept in the finished text, it is interesting because it includes the correct form *dominated*, which suggests that *dominted*, used in the final version of the sentence, is probably only a typo, since the learner is clearly able to spell the word correctly. What has not been represented in (2), but is visible in the PROCEED data, is the fact that the learner has paused on several occasions while typing this sentence. For example, in (2u), there is a long pause of 23 seconds just after *that*, which may be indicative of the learner’s difficulty in finishing the sentence. There is also a seven-second pause before the insertion of the *s*-letter at the end of *development* (2ab), which seems to correspond to a reviewing of the whole sentence, resulting in a last correction.

This example is an illustration of Murray’s (1980: 3) witty remark that “process can not be inferred from product any more than a pig can be inferred from a sausage.” It also points to the importance of considering the writing process next to the written product. Indeed, there have been calls in the literature to pay attention to the writing process. Back in the 1980s, Hairston (1982: 84) thus claimed that

we have to try to understand what goes on during the act of writing [...] if we want to affect its outcome. We have to do the hard thing, examine the intangible process, rather than the easy thing, evaluate the tangible product.

The use of computers as well as technologies like screencasting (recording of the screen activity) and keylogging (recording of the keys struck on the keyboard) have made the intangible more tangible: it is now possible to see the writing process unfold before our eyes, with all its deletions, insertions, substitutions, pauses, etc. Several recent studies have relied on such information to approach writing and have demonstrated its

usefulness for descriptive, theoretical or pedagogical purposes (see, e.g., Cislaru 2015; Lindgren and Sullivan 2019; Révész and Michel 2019).

Among the studies that use writing process data, the setting tends to be experimental, with data being collected specifically for this particular study, often among a small group of participants. In Breuer (2019), for example, the keystroke log files produced by 10 German students writing three texts in English and two in German are used to investigate the students' fluency in L1 (mother tongue) and L2 writing, revealing a higher degree of fluency in L1 than in L2 for most students. Sullivan and Lindgren (2002) test the pedagogical use of keystroke log files among four learners of English required to write a narrative text and demonstrate the positive effect of observing one's own composing process. In Elola and Mikulski (2016), a comparison is drawn between the screen activity of six learners of Spanish as a foreign language and 12 learners of Spanish as a heritage language, which brings to light similarities between the two groups (e.g. transfer of writing processes from the L1) as well as differences (e.g. more surface revisions but fewer meaning revisions in Spanish as a foreign language). The term *corpus* is hardly ever used in such studies, which may suggest that the data are not meant as a durable and reusable resource. A notable exception is Wengelin (2006), who describes her data sets, consisting of keystroke log files for Swedish texts, as corpora. Moreover, she shows how the techniques of corpus linguistics can be applied to the study of pauses in writing by looking for 'micro-contexts' made up of a pause preceded and followed by certain elements (e.g. a pause preceded by a typed letter and followed by a deletion). Cislaru and Olive (2018) similarly refer to their process data (different versions of texts in French, together with the keystroke log files) as a corpus. In addition, they explicitly mention corpus linguistics as one of the frameworks they draw inspiration from. Hamel and Séror (2016: 156) also use the term *corpus* to describe a collection of screencast videos showing the writing process of L2 learners of French and English. They point out that

such corpora represent new and exciting forms of empirical data which, once anonymized, could contribute to learner corpus projects that might be shared with others.

The *Process Corpus of English in Education* (PROCEED), as its name indicates, was designed as a corpus right from the start, meant as a durable and reusable resource bringing together a substantial amount of data supposed to be representative of a larger population. It relies on both keylogging and screencasting. It also comes with rich

metadata and comparable data in the learners' L1. The resource is described in more detail in Section 2, while Section 3 provides an overview of some of the research perspectives that the corpus offers. Section 4 concludes the article.

2. THE CORPUS

2.1. *A project in learner corpus research*

PROCEED can be described as a new type of learner corpus in the typology of learner corpora (cf. Gilquin 2015), namely a 'process learner corpus', which shows the process through which a text is composed on computer by language learners. It makes the writing process visible through keylogging and screencasting, two complementary methods to record the activity of writing a text on computer. The corpus aims to contribute to learner corpus research by providing a resource that allows for a novel and fine-grained approach to written performance, in the original sense of 'performance', that is, the *process* of doing something (in this case, writing a text).

The corpus project started in February 2017 with the collection of writing process data among a group of higher intermediate to advanced, mostly French-speaking students majoring in English at the University of Louvain (Belgium). Since then, additional data have been collected at least once a year among a new cohort of students each year. This is seen as the first step towards setting up an international project that seeks to collect similar data in other countries, among learners of English with different mother tongue backgrounds.

2.2. *The data*

Like traditional written learner corpora, PROCEED includes texts written by learners. These learner texts are written in English and are of the argumentative type, as this genre is thought to involve more complex writing processes than other text types like narrative texts (cf. Roca de Larios *et al.* 2002). Each learner begins by choosing a topic or quote among several options offered. They then have about 45 minutes to write a text of approximately 350 words defending their point of view. They are allowed to use online reference tools but are asked not to draw on secondary sources. These texts represent the written product.

In addition to the written product, the corpus includes writing process data. With the learners' permission, the keys struck on the keyboard are recorded by means of *Inputlog* (Leijten and Van Waes 2013) and the screen activity is recorded by means of *OBS Studio* during the whole writing task.² The *Inputlog* data take the form of log files, one per text, representing the different actions performed (letters typed, deletion, capitalisation, mouse click or movement, pauses, transition between Word document and other windows, etc.). These files can serve as a basis to carry out different types of analyses and to compute various statistics within *Inputlog*, e.g. linear analysis (with one action per line), revision matrix (a list of all the revisions), writing time, pausing time or number of revisions. Since they involve textual/numerical data, they can be searched by means of the techniques of corpus linguistics, although with adapted queries (cf. Wengelin's (2006) study, mentioned in Section 1). *Inputlog* has a replay function, which makes it possible to reconstruct the writing process in a video-like manner on the basis of the stored data. However, the function comes with a warning that an error-free replay of the process files cannot be guaranteed and with a recommendation for researchers relying on replay to resort to screencasting.

Screencasting with *OBS Studio* produces a faithful representation of the screen activity during the writing task. The *OBS Studio* data take the form of screencast videos, one per text. The videos can be navigated easily, and played at different speeds, using any multimedia player. While videos as such cannot be queried directly with the usual tools and techniques of corpus linguistics, they may be amenable to queries via alignment with the keystroke log files or via annotation. The *OBS Studio* videos can be aligned with the *Inputlog* data thanks to the video timeline and the timestamps associated with each action in *Inputlog*. A search on the textual data from *Inputlog* with text retrieval software could therefore generate hits from the *Inputlog* file that are linked to the corresponding part of the video. Annotation is another way of querying the screencast videos. A program like *ELAN* (Wittenburg *et al.* 2006) makes it possible to annotate videos with written information that describes their contents, by inserting annotation tiers which include attributes assigned to specific video segments (e.g. segments without any typing or involving the use of an online dictionary; see Laporte and Gilquin 2018 for an illustration). The information provided in the annotation can then be searched by means of text retrieval software.

² <https://obsproject.com> (12 March, 2021.)

Although PROCEED is first and foremost a learner corpus, consisting of non-native data produced by language learners, it was deemed relevant to include L1 data representing the learners' writing process in their mother tongue. This is because writing processes are said to display "conspicuous individual differences" (Sasaki 2000: 262), which may partly be the result of idiosyncratic behaviours that are language-independent, and hence valid regardless of whether the writer is writing in their L1 or in an L2. Comparing writers' behaviours in L1 and L2 is not only intrinsically interesting (cf. Thorson 2000; Stevenson *et al.* 2006), but it can also help distinguish these language-independent features from those that are due to the non-native nature of the writing process. The L1 data are collected according to the same principles as the L2 data: the learners have about 45 minutes to write a 350-word argumentative text on one of several set topics/quotes, while their screen and keyboard activity is recorded with their permission.

2.3. *The metadata*

As is the case with most learner corpora, PROCEED comes with rich metadata describing learners' profiles and collected via a questionnaire to be filled in by each participant. It includes personal information (age, gender, nationality, country of residence, etc.) as well as information about the learner's use and knowledge of languages (native language, parents' native languages, language(s) used in everyday life, language(s) of instruction at school, knowledge of foreign languages, etc.). Particular attention is paid to learners' exposure to English (number of years of English at school/university, proportion of classes taught in English, time spent in an English-speaking country, varieties of English they have been exposed to, etc.) as well as the kind of contexts in which they use English (estimation of the time spent doing certain activities, such as reading, watching TV or doing homework, in English). Learners are also asked to evaluate their (speaking, writing, listening, reading, pronunciation, grammar and spelling) skills in English. This comes as a complement to their score on the *LexTALE* vocabulary test, which has been shown to correlate with general tests of English proficiency (Lemhöfer and Broersma 2012). Finally, the questionnaire includes a few questions that are specifically related to the kind of corpus data collected, such as the type of keyboard learners usually use or whether they have been diagnosed with dyslexia.

Because typing speed is essential when considering aspects of the writing process such as fluency, learners are required to carry out a copy task, both in English and in their L1. The copy task was designed by the developers of *Inputlog*, within which the results of the task can be analysed. It can be done online, with the output file being directly downloadable from the website.³ It involves several activities: pressing two keys one after the other as quickly as possible, copying a sentence as many times as possible, copying combinations of three words and copying blocks of consonants.

The analysis of writing process data can provide insights into more cognitive aspects of language performance (cf. Section 3.1). For this reason, the PROCEED metadata also include measures of learners' cognitive abilities, which can be related to the writing process data and possibly account for some of the individual variation. These measures are collected by subjecting the participants to a battery of tests. Learners' verbal aptitudes (including vocabulary learning and grammatical inferencing) are tested through some of the *LLAMA* (language-independent) tests (Meara and Rogers 2019). Their non-verbal aptitudes are tested by means of *Raven's Matrices* (Raven and Raven 2003), which measure abstract reasoning (fluid intelligence). In addition, the *Psychology Experiment Building Language* (PEBL) interface (Mueller 2012) is used to assess working memory capacity (by means of the *Operation Span* task; cf. Hegarty and Dufflecoat Enterprises 2014) as well as response inhibition and interference suppression (by means of the *Flanker* and *Simon* tasks; cf. Mueller 2011a, 2011b).

3. RESEARCH AND PEDAGOGICAL PERSPECTIVES

3.1. Writing process research

Besides the kind of research that is traditionally possible on the basis of written learner corpora, the PROCEED data have great potential for research into the writing process. By combining keylogging and screencasting, they present an accurate picture of the way learners of English compose their texts, with unprecedented detail on the actual mechanics of the process. This information can be used for descriptive, explanatory and theoretical purposes.

In terms of description, the keylogging data provide comprehensive statistics about aspects that have to do with writing fluency (number, duration and location of

³ <http://inputlog.ua.ac.be/Website/copytask/tasks.html> (8 March, 2021.)

pauses, type and number of revisions, etc.). As against the conventional approaches that measure fluency as the number of words produced overall or the mean number of words produced per minute (cf. Sasaki 2004), the keylogging-based approach considers writing fluency in its multidimensionality (cf. Van Waes and Leijten 2015). This focus on the notion of fluency also opens up new possibilities for comparing learner writing and speech. In addition, keylogging and screencasting data make it possible to examine the use of online resources during the writing process, such as secondary sources (Leijten *et al.* 2019) or writing tools (Gilquin and Laporte forthcoming, based on the annotation of PROCEED videos with *ELAN*). The data could also be used to carry out a dynamic discourse analysis, looking at how discourse is created in real time (e.g. paragraph formation, development of rhetorical functions) or what strategies learners adopt to compose a text (e.g. linear composition or outline that is progressively fleshed out).

A further use of PROCEED is for explanatory purposes. The writing process data can help account for the origin of certain features of the finished texts. A lack of tense agreement between main clause and subclause, for example, may turn out to be due to the fact that the tense of the main verb was changed at some stage but the writer failed to adapt the tense of the verb in the subclause (cf. Gilquin 2021). The data can also help uncover more cognitive aspects of writing performance (cf. Spelman Miller *et al.* 2008). Revisions may thus point to a lack of automaticity for certain language components (e.g. the subject-verb agreement rule, if the verb form regularly needs to be revised) or to phenomena of avoidance (e.g. avoidance of the passive, if passive structures are systematically aborted), which are typically very difficult to discover on the basis of written texts only. Seeing what words are produced together in one go (the so-called ‘bursts’, see Chenoweth and Hayes 2001) can also give an indication of the constructions that are stored as wholes in the mind (Gilquin 2020).

From a theoretical perspective, writing process data such as those found in PROCEED can help develop or improve models of writing, as shown in Leijten *et al.* (2014) with keylogging data. The design of PROCEED, consisting of texts produced by the same writers in their L1 and in L2 English, could lead to the development of bilingual writing models representing native and non-native writing, and showing how L1 and L2 writing abilities interact with each other. The metadata associated with each writer might even make it possible to adapt a general writing model to individual

variation, most notably through the empirical measures of working memory, which is part and parcel of most writing models (cf. Kellogg 1996; Hayes 2012).

3.2. *Teaching and testing applications*

Next to its use for research purposes, PROCEED also has potential applications for teaching and testing. The most immediate pedagogical application is to use PROCEED as a local learner corpus, that is, a corpus that is collected by the teacher among—and for the benefit of—his or her own students (Seidlhofer 2002). In other words, the learners are both contributors to and users of PROCEED. After collecting data from a group of learners, they can each be given access to their screencast video and be required to watch (part of) it, so as to become aware of how they actually compose a text. Additionally, clips from some learners' videos can be selected and shown to the members of the group, to illustrate effective strategies that could be useful to them (e.g. highlighting words to be checked later in a dictionary, so that the flow of ideas does not get interrupted). Learners can also be presented with some statistics describing their writing behaviour. On the basis of a keystroke log file, *Inputlog* can generate a user report that summarises some important facts about the user's writing process, such as the time they have been writing vs. pausing or the number of revisions they have made (Vandermeulen *et al.* 2020). The report also includes a graph representing the writing process which, with some explanations, could help learners visualise their own writing behaviour, and possibly compare it with the behaviour of other learners in the group or that of native writers (see Gilquin 2019 for a pedagogical intervention based on PROCEED as a local learner corpus). The PROCEED data can also be used as pedagogical materials for learners other than those among whom the data were collected. Video clips illustrating different writing strategies (effective or less effective) could be shown to learners to help them reflect on the act of writing and how best to compose a text. The process graphs generated by *Inputlog* could also be used as a basis to exemplify various writing behaviours (e.g. revising the text as one goes along or leaving some time at the end to revise the whole of it).

The writing process data from PROCEED can also serve testing purposes. While the testing of writing skills typically only relies on the quality assessment of the finished text, considering the writing process too could result in a more fine-grained evaluation of writing performance (cf. Ranalli *et al.* 2018). Thus, it would make sense, as is the

case for speech, to include a criterion like writing fluency, which would aim to assess how smooth the writing process is. The PROCEED data, and in particular the analysis of the keystroke log files, could provide the necessary statistics to empirically assess the writing fluency of the learners who contributed to the corpus. Their writing fluency in the mother tongue could even be taken into account to provide a tailor-made yardstick for each learner. Another aspect that could be relevant to the evaluation of writing skills is consultation behaviour, that is, the way in which learners resort to online writing tools like dictionaries or thesauri, as using these tools effectively may be seen as an important component of writing performance. Again, this can be examined empirically for the contributors to the corpus, using the screencast videos. The analysis of such aspects of the writing process in PROCEED could also help improve writing assessment on a more general level, for other learners than those who contributed to the corpus. By bringing together data from a large number of participants, PROCEED can be said to be representative of a certain population of learners. It can therefore be exploited to determine the typical writing behaviour of learners at a given proficiency level, for example in terms of pausing time or number of revisions, and to set this as the expected standard. Other learners with a similar profile can then be evaluated against this corpus-derived standard.

4. CONCLUSION

This article has introduced a new resource, PROCEED, which also represents a new type of corpus to investigate learner writing. Its unique combination of written texts, screencast videos, keystroke log files, rich metadata including cognitive measures, and equivalent L1 data offers an unparalleled opportunity to study the process through which learners write texts. It also opens new perspectives in terms of research and applications: study of writing fluency and comparison with spoken fluency; analysis of learners' use of online writing tools; dynamic discourse analysis taking the development of discourse into account; exploration of cognitive aspects of writing performance; theoretical modelling of the bilingual writing process; pedagogical interventions involving learners' examination of their own writing behaviour; addition of a 'process' component to the assessment of writing skills, based on corpus-derived standards; etc.

While collecting and analysing corpus data of the PROCEED type implies different routines than those followed in traditional learner corpus research, this

description of the PROCEED project will hopefully have demonstrated the value of what could be referred to as ‘process learner corpus research’, and the significance of its possible applications. The potential of PROCEED will arguably continue to increase as the corpus keeps growing in size and in diversity of learner profiles.

REFERENCES

- Breuer, Esther Odilia. 2019. Fluency in L1 and FL writing: An analysis of planning, essay writing and final revision. In Eva Lindgren and Kirk P. H. Sullivan eds. *Observing Writing: Insights from Keystroke Logging and Handwriting*. Leiden: Brill, 190–211.
- Centre for English Corpus Linguistics. 2020. Learner corpora around the world. Louvain-la-Neuve: Université catholique de Louvain. <https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html> (23 December, 2020.)
- Chenoweth, N. Ann and John R. Hayes. 2001. Fluency in writing: Generating text in L1 and L2. *Written Communication* 18/1: 80–98.
- Cislaru, Georgeta ed. 2015. *Writing(s) at the Crossroads: The Process–Product Interface*. Amsterdam: John Benjamins.
- Cislaru, Georgeta and Thierry Olive. 2018. *Le processus de textualisation. Analyse des unités linguistiques de performance écrite*. Louvain-la-Neuve: De Boeck Supérieur.
- Elola, Idoia and Ariana M. Mikulski. 2016. Similar and/or different writing processes? A study of Spanish foreign language and heritage language learners. *Hispania* 99/1: 87–102.
- Gilquin, Gaëtanelle. 2015. From design to collection of learner corpora. In Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier eds. *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 9–34.
- Gilquin, Gaëtanelle. 2019. Screencasting and keylogging as pedagogical tools to enhance writing skill development. Paper presented at *EUROCALL 2019, Louvain-la-Neuve, Belgium, 28–31 August 2019*.
- Gilquin, Gaëtanelle. 2020. In search of constructions in writing process data. *Belgian Journal of Linguistics* 34: 99–109.
- Gilquin, Gaëtanelle. 2021. Hic sunt dracones: Exploring some *terra incognita* in learner corpus research. In Anna Čermáková and Marketa Malá eds. *Variation in Time and Space: Observing the World through Corpora*. Berlin: De Gruyter, 65–86.
- Gilquin, Gaëtanelle and Samantha Laporte. Forthcoming. The use of online writing tools by learners of English: Evidence from a process corpus. *International Journal of Lexicography*.
- Hairston, Maxine. 1982. The winds of change: Thomas Kuhn and the revolution in the teaching of writing. *College Composition and Communication* 33/1: 76–88.
- Hamel, Marie-Josée and Jérémie Séror. 2016. Video screen capture to document and scaffold the L2 writing process. In Catherine Caws and Marie-Josée Hamel eds. *Language-Learner Computer Interactions: Theory, Methodology and CALL Applications*. Amsterdam: John Benjamins, 137–162.
- Hayes, John R. 2012. Modeling and remodeling writing. *Written Communication* 29/3: 369–388.

- Hegarty, David L. and Dufflecoat Enterprises. 2014. *The PEBL Operation Span Task*. <http://pebl.sourceforge.net/battery.html> (12 March, 2021.)
- Kellogg, Ronald T. 1996. A model of working memory in writing. In C. Michael Levy and Sarah Ransdell eds. *The Science of Writing: Theories, Methods, Individual Differences, and Applications*. Mahwah, NJ: Erlbaum, 57–71.
- Laporte, Samantha and Gaëtanelle Gilquin. 2018. Annotating the use of online writing resources in a video corpus of written process data in ELAN. Annotation manual version 1.1. CECL Papers 2. Louvain-la-Neuve: Université catholique de Louvain. <https://uclouvain.be/en/research-institutes/ilc/cecl/cecl-papers.html> (8 March, 2021.)
- Leijten, Mariëlle and Luuk Van Waes. 2013. Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication* 30/3: 358–392.
- Leijten, Mariëlle, Luuk Van Waes, Iris Schrijver, Sarah Bernolet and Lieve Vangehuchten. 2019. Mapping master's students' use of external sources in source-based writing in L1 and L2. *Studies in Second Language Acquisition* 41/3: 555–582.
- Leijten, Mariëlle, Luuk Van Waes, Karen Schriver and John R. Hayes. 2014. Writing in the workplace: Constructing documents using multiple digital sources. *Journal of Writing Research* 5/3: 285–337.
- Lemhöfer, Kristin and Mirjam Broersma. 2012. Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods* 44/2: 325–343.
- Lindgren, Eva and Kirk P. H. Sullivan eds. 2019. *Observing Writing: Insights from Keystroke Logging and Handwriting*. Leiden: Brill.
- Meara, Paul M. and Vivienne E. Rogers. 2019. *The LLAMA Tests v3*. Cardiff: Lognostics.
- Mueller, Shane T. 2011a. *The PEBL Flanker Task*. <http://pebl.sourceforge.net/battery.html> (12 March, 2021.)
- Mueller, Shane T. 2011b. *The PEBL Simon Interference Task*. <http://pebl.sourceforge.net/battery.html> (12 March, 2021.)
- Mueller, Shane T. 2012. The Psychology Experiment Building Language, Version 0.13. <http://pebl.sourceforge.net> (23 December, 2020.)
- Murray, Donald M. 1980. Writing as process: How writing finds its own meaning. In Timothy R. Donovan and Ben W. McClelland eds. *Eight Approaches to Teaching Composition*. Urbana, IL: National Council of Teachers of English, 3–20.
- Ranalli, Jim, Hui-Hsien Feng and Evgeny Chukharev-Hudilainen. 2018. Exploring the potential of process-tracing technologies to support assessment for learning of L2 writing. *Assessing Writing* 36: 77–89.
- Raven, John and Jean Raven. 2003. Raven Progressive Matrices. In R. Steve McCallum ed. *Handbook of Nonverbal Assessment*. Boston, MA: Springer, 223–237.
- Révész, Andrea and Marije Michel. 2019. State of the scholarship: Introduction. *Studies in Second Language Acquisition* 41/3: 491–501.
- Roca de Larios, Julio, Liz Murphy and Javier Marín. 2002. A critical examination of L2 writing process research. In Sarah Ransdell and Marie-Laure Barbier eds. *New Directions for Research in L2 Writing*. Dordrecht: Kluwer Academic Publishers, 11–47.
- Sasaki, Miyuki. 2000. Toward an empirical model of EFL writing processes: An exploratory study. *Journal of Second Language Writing* 9/3: 259–291.

- Sasaki, Miyuki. 2004. A multiple-data analysis of the 3.5-year development of EFL student writers. *Language Learning* 54/3: 525–582.
- Seidlhofer, Barbara. 2002. Pedagogy and local learner corpora: Working with learning-driven data. In Sylviane Granger, Joseph Hung and Stephanie Petch-Tyson eds. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins, 213–234.
- Spelman Miller, Kristyan, Eva Lindgren and Kirk P. H. Sullivan. 2008. The psycholinguistic dimension in second language writing: Opportunities for research and pedagogy using computer keystroke logging. *TESOL Quarterly* 42/3: 433–454.
- Stevenson, Marie, Rob Schoonen and Kees de Glopper. 2006. Revising in two languages: A multi-dimensional comparison of online writing revisions in L1 and FL. *Journal of Second Language Writing* 15/3: 201–233.
- Sullivan, Kirk and Eva Lindgren. 2002. Self-assessment in autonomous computer-aided second language writing. *ELT Journal* 56/3: 258–266.
- Thorson, Helga. 2000. Using the computer to compare foreign and native language writing processes: A statistical and case study approach. *Modern Language Journal* 84/2: 155–170.
- Vandermeulen, Nina, Mariëlle Leijten and Luuk Van Waes. 2020. Reporting writing process feedback in the classroom: Using keystroke logging data to reflect on writing processes. *Journal of Writing Research* 12/1: 109–139.
- Van Waes, Luuk and Mariëlle Leijten. 2015. Fluency in writing: A multidimensional perspective on writing fluency applied to L1 and L2. *Computers and Composition* 38/A: 79–95.
- Wengelin, Åsa. 2006. Examining pauses in writing: Theory, methods and empirical data. In Kirk P. H. Sullivan and Eva Lindgren eds. *Computer Keystroke Logging and Writing: Methods and Applications*. Amsterdam: Elsevier, 107–130.
- Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*, 1556–1559.

Corresponding author

Gaëtanelle Gilquin

Université catholique de Louvain

Collège Erasme

Place Blaise Pascal 1, bte L3.03.33

B-1348 Louvain-la-Neuve

Belgium

Email: gaetanelle.gilquin@uclouvain.be

received: December 2020

accepted: March 2021

published online: April 2021

The compilation of a developmental spoken English corpus of Turkish EFL learners

Ece Genç-Yöntem – Evrim Eveyik-Aydın
Yeditepe University / Turkey

Abstract – Although compiling a spoken learner corpus is not a recent enterprise, the number of developmental learner spoken corpora in the field of corpus linguistics is not satisfactory. This report describes the compilation of the *Yeditepe Spoken Corpus of Learner English* (YESCOLE), a 119,787-word corpus of Turkish students' spoken English at tertiary level. YESCOLE was compiled to generate a developmental corpus of spoken interlanguage by collecting samples from learners of different English proficiency levels at regular short intervals over seven months. In order to shed light on the laborious methodology of compiling the developmental spoken learner corpus, this paper elucidates the steps taken to build YESCOLE and discusses its potential benefits for research and instructional purposes.

Keywords – learner corpus; spoken corpus; corpus compilation; developmental corpus; EFL

1. INTRODUCTION

Data collection timing has been an important criterion in learner corpus research. In learner corpora studies, data can be collected either at one point in time or repeatedly over time depending on the purpose of the research study. The former are called synchronic corpora and the latter are diachronic corpora (Gilquin 2015: 13). Most of the learner corpora are synchronic and have a cross-sectional design (e.g. the *International Corpus of Crosslinguistic Interlanguage*, ICCI; Tono and Díez-Bedmar 2014). Since it is difficult to follow the same learner or group of learners over time, compiling diachronic corpora (e.g. longitudinal and developmental corpora) is quite challenging for many researchers. An example to a longitudinal learner corpus is the *Longitudinal Database of Learner English* (LONGDALE) in which learners were tracked over three years by collecting data once a year (Meunier 2016). Some learner corpora are called developmental when the data are collected more densely. In such corpora, “learner performance is documented at close intervals or at all points of production” (Belz and Vyatkina 2008: 33). In the study

by Belz and Vyatkina (2008), the research corpus, *Telecollaborative Learner Corpus of English and German* (Telekorp), was defined as developmental on the grounds that learners were followed over a two-month period. In this vein, both longitudinal and developmental corpora are rich sources that display progress of learners (Gilquin 2015: 13).

The present paper reports on the compilation steps of the *Yeditepe Spoken Corpus of Learner English* (YESCOLE), which is also a developmental corpus that has the potential to fill a significant gap in the field for a number of reasons. First of all, according to the information obtained from the *Centre for English Corpus Linguistics* (CECL) at Université catholique de Louvain,¹ it should be indicated that, except for some corpora containing samples of learners' spoken production (e.g. the *Louvain International Database of Spoken English Interlanguage*, LINDSEI),² the existing learner corpora in the field are in the written mode. Secondly, while existing corpora include language samples from learners with different first language (L1) backgrounds, the number of spoken corpora that involve language produced by L1 Turkish learners of English is found to be rather limited, as shown in Table 1.

Corpus Name	L1	Mode	Size in words
The Turkish component of the LINDSEI (LINDSEI-TR)	Turkish	Spoken	80,813
<i>Corpus of Learner Monologues</i> (CLM)	Turkish	Spoken	6,151
<i>Turkish Corpus of Spoken Learner English</i> (TC-SLE)	Turkish	Spoken	1,500 (in progress)

Table 1: List of spoken learner corpora in which Turkish is L1 and English is the target

Among the few existing corpora, the Turkish component of the LINDSEI (LINDSEI-TR) was compiled by Kilimci (2014), and it includes almost 50 advanced level English learners' interviews each lasting about 15 minutes. The tasks within the interviews range from telling a story to answering a question and describing a picture. Another spoken learner corpus of Turkish EFL learners, the *Corpus of Learner Monologues* (CLM), was compiled by Demirel and Şahin (2015) to identify lexical problems in spoken English, and it comprises 35 participants' two-minute talks on selected *International English Language Testing System* (IELTS)³ speaking topics. The proficiency levels of the English learners contributing to CLM range from intermediate to upper-intermediate. The last corpus, the *Turkish Corpus of Spoken Learner English* (TC-SLE), was compiled by

¹ <https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html>

² <https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html>

³ <https://www.ielts.org/>

Demirel and Kazazoğlu (2015), and it consists of two-minute talks of Turkish learners of English whose levels of proficiency range from intermediate to advanced on two selected IELTS speaking topics. TC-SLE is still in progress and intends to include learners' monologues and language use in classroom contexts and group work activities.

Despite their merits, the above-mentioned spoken corpora, compiled in a Turkish EFL context, do not comprise longitudinal data from the same learner or learners over time. As stated previously, one of the main reasons for the scarcity of longitudinal corpora is that it is rather challenging for researchers to track the same learner or group of learners over time. Moreover, many researchers face data attrition issues in a way that the same participant stops contributing to the corpus over time, which discourage them to conduct studies that will take time. Therefore, not only does a low number of scholars find opportunities to collect learner data longitudinally, but also such learner corpora in the field are mostly in the written mode. However, developmental and longitudinal learner corpora can shed light on the changes in interlanguage and difficulties that are experienced during the course of language learning. This was also pointed out by Thewissen (2013), who analyzed the error-tagged quasi-longitudinal corpus of 223 essays obtained from the *International Corpus of Learner English* (Granger *et al.* 2009) and demonstrated the accuracy development of B1, B2, C1 and C2 learners of English by examining the corpus-driven learner errors. This implies that there exists an ever-increasing need for developmental and longitudinal learner spoken corpora in foreign language contexts.

In view of this, we have compiled YESCOLE, a specialized spoken corpus of L1 Turkish learners of English as part of a doctoral study to investigate the spoken interlanguage of English-major students who attend a language preparatory program at Yeditepe University, a foundation university in Istanbul, Turkey. This paper covers a) in Section 2, the steps taken to compile YESCOLE, and b) in Section 3, a discussion of the research potential that such a learner spoken corpus offers. Hence, not only does the paper present a guiding methodology for those who pursue studies in a similar vein, but it also highlights how such corpora enable studies on learners' interlanguage.

2. THE COMPILATION OF YESCOLE

2.1. Representativeness and descriptive features

A learner corpus should be large and representative enough to address the target research questions (Granger 2004: 125). The design criteria (e.g. population, proficiency level, tasks, mode and timing) should be selected carefully before the corpus compilation. As highlighted by Rea Rizzo (2010: 3), the corpus can also be specialized when data are collected from specific groups (e.g. a spoken learner corpus) or genres (e.g. a corpus of English as a Lingua Franca in academic settings) and when it is aligned with corpus builders' own research purposes (e.g. studying the changes in language use of Spanish learners of English). Specialized corpora have lately become a preferred way to answer specific research questions, especially in EFL contexts.

Considering this, as displayed in Table 2, YESCOLE is a specialized corpus that includes spoken performance of young adult Turkish EFL learners at Yeditepe University, Turkey. It was compiled to generate a developmental corpus of spoken interlanguage by collecting spoken samples from learners at three different levels of English proficiency (A2, B1 and B2 according to the *Common European Framework of Reference for Languages*, CEFR)⁴ to investigate grammatical and lexical errors over time in learners of English. YESCOLE represents the academic genre because the data collection setting is a university preparatory school, and data were elicited from the learners as part of their oral exams. The oral exams, which include tasks that require learners to make speeches to discuss the causes/effects or advantages/disadvantages of something, are used quite often in that prep school context. Such oral exams are typical examples of classroom genres that represent spoken academic discourse.

Type	Specialized corpus / learner corpus/ developmental corpus
Mode	Spoken
Population	Young adult English-major Turkish EFL learners at a foundation university preparatory school in Turkey
English proficiency level	A2 (pre-intermediate), B1(intermediate) and B2 (upper-intermediate)
Genre	Academic/ Oral exam/ (monologue/ non-interactive)

Table 2: Features of YESCOLE

YESCOLE comprises four sub-corpora: 1) YESCOLE-A2 (corpus of pre-intermediate level Turkish EFL learners), 2) YESCOLE-B1 (corpus of intermediate level Turkish EFL learners), 3) YESCOLE-B2 (corpus of upper-intermediate level Turkish EFL learners),

⁴ See CEFR manual (Council of Europe 2005) for detailed information regarding each level of language proficiency.

and 4) YESCOLE-LONG (developmental corpus of A2, B1 and B2 level Turkish EFL learners). The corpus is not tagged by part-of-speech (POS); however, it is error-tagged (see 2.3.6.). The details related to the spoken corpus such as total tokens, types and utterance counts were computed using *AntConc* 3.5.8 (Anthony 2019). These are shown in Table 3.

	YESCOLE-A2	YESCOLE-B1	YESCOLE-B2	YESCOLE-LONG	YESCOLE-Total
Total word	13,806	50,571	19,088	36,322	119,787
Total type	1,462	3,066	1,901	2,053	3,922

Table 3: Size of YESCOLE and its sub-corpora

In total, YESCOLE comprises 119,787 words. YESCOLE-B1 is the largest in size (50,571 words), and it is followed by YESCOLE-LONG (36,322 words) and YESCOLE-B2 (19,088 words).

2.2. Participants

The spoken data used to generate the specialized spoken corpus of learner English were collected through convenient sampling from 105 Turkish young adult EFL learners who study in the English language preparatory program specifically intended for the students of Translation Studies (TRA), English Language and Literature (ELIT) and English Language Teaching (ELT) at Yeditepe University. Participation was voluntary; therefore, a consent form requesting students' permission to allow their instructors to record their speech during exams was prepared. 105 learners out of 112 granted permission and filled in a learner profile questionnaire, which provided demographic information related to age, gender, language background (e.g. their L1 and when they started learning English), and whether they had any health problems (e.g. hearing or speaking impairment or learning disability). The demographic data revealed that the learners' average age was 19. Out of 105, 80 participants were female and 25 were male. The average age at which they started learning English as a foreign language was nine. They were all L1 speakers of Turkish and none of the students had health problems.

2.3. Steps taken to compile YESCOLE

2.3.1. Checking the proficiency levels of the participants

Since the initial purpose of collecting learner English spoken data was to examine the learners' progress over time, the participants were placed into three groups on the basis of the *Oxford Quick Placement Test* (OQPT) (2001), which can be used to provide information about the proficiency level of learners. The OQPT also offers a chart of equivalent levels to be interpreted with respect to the levels of the CEFR (2005). Table 4 summarizes the number, gender, and proficiency level of learners in the corpus.

Gender	A2 level	B1 level	B2 level	Total
Female	19	44	17	80
Male	3	12	10	25
Total	22	56	27	105

Table 4: Number of participants according to gender and proficiency levels

According to the classification offered by OQPT, participants who received a score between 18 and 29 out of 60 were placed into A2; those who received a score between 30 and 39 were placed into B1; and those whose score ranged from 40 to 47 were placed into B2. These results indicated that 27 participants were B2 level (upper-intermediate), 56 were B1 level (intermediate), and 22 were A2 level (pre-intermediate).

2.3.2. Selection of the prompts to elicit oral data

Different techniques to elicit L2 oral data have been reported in the literature. Some of these ways include learner monologues on a given topic (e.g. Demirel and Şahin 2015; Yıldız 2016), tasks of oral argumentation (e.g. Kormos and Dörnyei 2004), picture description (e.g. Yuan and Ellis 2003; Ellis and Yuan 2004), role-plays (e.g. Ting *et al.* 2010; de Jong *et al.* 2013), oral interviews (e.g. Boers *et al.* 2006; Huang 2011), story-telling (e.g. Khan 2011) and course presentations (e.g. Aşık and Cephe 2013). Although different tasks require different cognitive demands (Skehan 1998), asking for oral argumentation was acknowledged to be an effective way to elicit L2 learners' speech in the literature (e.g. Masrom *et al.* 2015). Therefore, speaking prompts that elicit students' opinions on different topics were used to compile YESCOLE because a) the students' performance is assessed with these questions in the program, and b) the oral elicitation

technique should be similar to oral argumentation so as to benefit from the advantages that this task brings with it (e.g. a substantial amount of spoken performance data).

Thus, the corpus consists of Turkish EFL learners' speaking test performance, which is based on their monologic talks on given speaking prompts during their oral exams. In accordance with the requirements of the prompt, the participants talk about causes or effects of a particular topic, offer solutions or suggestions for a particular problem, talk about advantages or disadvantages of a particular topic, and provide reasons for their opinion. These prompts elicit students' opinions on various topics discussed in their courses. Some prompts elicited from their course materials can be exemplified as:

(1a) Attendance at university should not be obligatory. To what extent do you agree or disagree with this idea? State your reasons.

(1b) Obesity has become the main concern for many young people. What are the effects of this situation?

2.3.3. Preparing the setting for oral recordings

After taking students' voluntary consent, the next step was to collect audio-recordings of spoken data. The English monologues on different topics of the 105 participants were audio-recorded in three speaking exams, including one quiz and two achievement exams during the 16-week semester. At the end of the semester, 29 participants from the initial 105 were enrolled in the summer school, so their spoken data were also recorded in three speaking exams during the 12-week summer semester. As can be seen in Figure 1, spoken data were collected at short, regular intervals from the same students to see their progress in line with the aim of the study. This longitudinal design was necessary to reach a rich source of data.

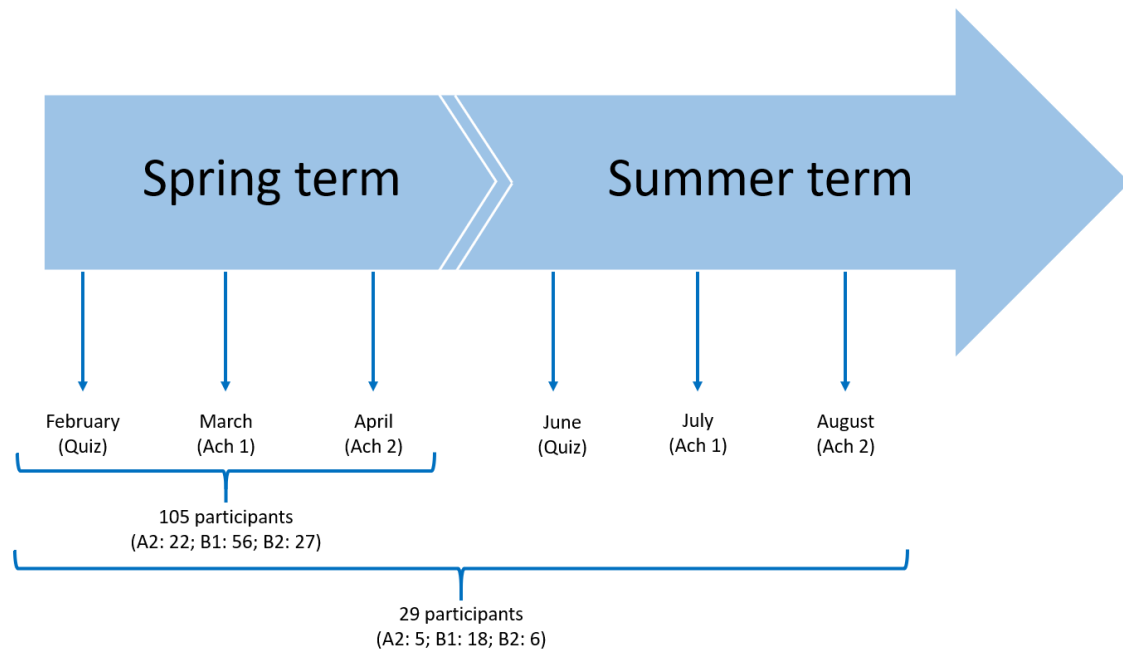


Figure 1: Timescale of recording spoken data

Speaking exams given to each participant were administered in private sessions in an exam room by two instructors and recorded with a professional quality audio-recorder. The students, who blind picked one of the speaking prompts on separate pieces of papers, were given one-minute planning time to think about their response. The importance of pre-task planning time was pointed out by Ahangari and Abdi (2011) in terms of its benefit to increase the quantity and complexity of the oral performance. Likewise, Skehan (1998) and Ortega (1999) highlighted the positive effects of planning time on L2 tasks, which balance the cognitive load and reduce speaking anxiety.

2.3.4. Transcription of audio-recordings

Approximately 20 hours of recording was done in the spring semester and five hours of recording was gathered during the summer school. After the recordings, the 402 sound files were checked for their quality and saved in the researcher's computer. Each participant was given a number, from one to 105, both to facilitate data storage and to keep data anonymous for ethical considerations. The audio files were also grouped according to the participants' proficiency levels and time of data collection. In order to code the files, the time of the data collection was indicated with a combination of the letter T and a number. In other words, T1, T2 and T3 were used to refer to the first, second and third data collection periods during the spring semester, respectively. T4, T5 and T6,

however, were used to indicate the fourth, fifth and sixth data collections in summer. Hence, to illustrate, the first recording of participant 9 at B2 proficiency level was coded as 9B2T1.

As the amount of spoken data to be transcribed was large, the workload of transcription was shared with a doctoral student in English Language Education. To transcribe the spoken data efficiently and consistently, a set of rules was followed. For example, as the focus in the corpus was on learner errors, spoken data were transcribed without correcting any error, false starts, repetitions and reformulations. To ensure consistency throughout spoken data transcription, the chat transcription conventions commonly used in the field (MacWhinney 2000) were adapted. To illustrate, fillers (e.g. *uh* and *uhm*) and unintelligible speech were marked with specific codes: [&-um] was used for fillers and [xxx] was used to transcribe unintelligible speech. Pauses were also indicated with [.] or [..] (if the duration is more than 7 seconds). The audio files were transcribed in *Microsoft Word* files. Then, these files were converted into plain text format in *Notepad* using *AntFileConverter* 1.2.1 (Anthony 2017) so that *AntConc* 3.5.8 (Anthony 2019) could be used for corpus description and analysis.

2.3.5. Identifying the utterances as units of analysis

In spoken learner language studies, one of the first decisions to be made is whether to count sentences or utterances as units of spoken language. In fact, researchers prefer the term ‘utterance’ or ‘C-unit’ (conversational unit) to refer to the basic unit of spoken language. Within the spoken corpus construction and analysis process, it is important to pinpoint utterances appropriately and consistently because analyses are conducted on the basis of utterances (Yaman *et al.* 2008). However, identifying utterances is very difficult in spoken production. There have been some techniques used to determine them, including intonation (Traum and Heeman 1997), probabilistic language models (Stolcke *et al.* 1998) and speech intervals as input units (Worm 1998).

In this study, the participants’ spoken production is monologic and academic, so it includes unscripted monologues. However, it also reflects the features of spoken language such as disfluencies, repetitions, retraces, and incomplete sentences. Before corpus-based linguistic data analysis, utterances were identified as they are the basic units of analysis

in this study (see Table 5 for utterance numbers). Both intonation and speech intervals were used as techniques to detect utterances in YESCOLE.

	YESCOLE- A2	YESCOLE- B1	YESCOLE- B2	YESCOLE- LONG	YESCOLE- Total
Utterance count	1,509	3,550	1,420	2,652	9,131

Table 5: Utterance numbers in YESCOLE

2.3.6. Corpus annotation and analyses

After the long transcription process, labels or tags were added to the corpus so that it can be automatically analyzed using a corpus analysis tool. This step is called ‘corpus annotation’ or ‘coding’. In this process, some information such as tags or labels is inserted into the original transcriptions. There are some tools that annotate the native corpus data automatically (e.g. SPPAS, Bigi 2015); however, due to the nature of learner language, some problems (e.g. inconsistent annotation) might occur in learner corpus annotation. Researchers can develop their own tags and coding schemes depending on their objectives. Tags are generally hand-coded on the transcripts via corpus analysis tools such as *AntConc* (Anthony 2019) and *Computerized Language Analysis* (CLAN) (MacWhinney 2000). Yet, reliability tests should be conducted in order to make sure that the data have been consistently tagged.

In accordance with the purpose of our research, YESCOLE was error-tagged by adding the tags to show grammatical and lexical errors in the data, and this is called ‘error annotation’. To describe the error types appropriately, Dulay *et al.*’s (1982) surface strategy taxonomy (omission, addition, misformation [misselection] and misordering of linguistic elements) was followed. To use a standard format for error tagging and to make it consistent, an annotation scheme was developed, and each error type was marked with a different tag. The tags for errors in YESCOLE were created by adapting those used in the CLAN manual (MacWhinney 2000). To illustrate, an error was identified with an asterisk in square brackets in the text after the error. For example, the omission of plural -s error in English was coded as: [*ms:a:0s]. In this code, **ms* stands for morpho-syntactic error, *a* indicates that it is an agreement error, and *0s* stands for omission of plural -s. In another example from the same linguistic level, addition of plural -s was coded as [*ms:a:+s]. In this code, *+* stands for addition of plural -s. Examples of an error-tagged utterance are given in Table 6.

Linguistic level	Error type	Example	Error Tag
Morpho-syntax	Omission of comparative <i>-er</i>	<i>He is short [*ms:0er] <shorter> than his sister.</i>	[*ms:0er]
Morpho-syntax	Addition of plural <i>-s</i>	<i>He gave me [*ms:a:+s] <advice>.</i>	[*ms:a:+s]

Table 6: Examples of an error-tagged utterance

Moreover, *AntConc* (Anthony 2019) was used to annotate corpus data by inserting tags or labels to facilitate word search and corpus analysis. These tags can be hidden or shown in the search results.

3. POTENTIAL BENEFITS OF YESCOLE FOR RESEARCH AND EFL INSTRUCTION

YESCOLE includes spoken samples of A2, B1 and B2 level (according to the CEFR) Turkish learners of English at tertiary level, and it will be expanded by collecting spoken data from A1 and C1 level learners at regular time intervals. After the spoken samples have been collected to describe the continuum of language proficiency, the corpus will be made available for researchers. The potential of YESCOLE for research and instruction is a good example for those willing to compile such developmental corpora, which are truly lacking in the field. According to McEnery and Gabrielatos (2006: 49), corpora have assisted language-related inquiry in four main aspects: 1) depiction of language and creation of reference materials; 2) lexicogrammatical analysis of language; 3) EFL instruction; and 4) noticing changes in a language. In addition, regarding spoken learner corpus building, Du Bois (1991: 73) states that

a transcription of spoken discourse can provide a broad array of information about these and other aspects of language, with powerful implications for grammar, semantics, pragmatics, cognition, social interaction, culture, and other domains that meet at the crossroads of discourse.

Considering this, compiling a specialized corpus of learners' developmental spoken English will offer many benefits and areas of application for research and EFL instruction.

Since YESCOLE is a developmental corpus, spoken data collected at different times reflect potential changes in learner English. As learner language has been claimed to have its own rules and developmental patterns (Selinker 1972), learner corpus research has been of great help to describe and track the progress across different language proficiency levels, observe the difficulties learners face in the process of learning and call for action (Thewissen 2013). One of the best ways to identify the problems which learners

face is to analyze the language used by them. By understanding the specific issues that a certain group of students from the same L1 language background face, researchers can more precisely understand and compare the different stages of that group's acquisition of English through the compilation of learner corpora in their own contexts. In spite of the difficulty in collecting spoken data repeatedly from a learner or a group of learners over time, developmental/longitudinal studies using corpus-approach to track EFL spoken performance should contribute more to the field.

De Cock (2010) highlighted that there is a need for studies using spoken learner corpora in the classroom. This holds especially true in the Turkish EFL context where there is a great need for diachronic or developmental/longitudinal spoken learner corpora to gain a better understanding of learner English. Such learner corpora can also be used in the classroom in order to support EFL instruction. As EFL learners find it challenging to speak accurately in English, learner corpora in spoken mode might show not only the weaknesses and the strengths of English language learners but also the progress of reducing language errors. Even specific language structures, such as the use of passive voice or adjectival clauses, can be investigated and specific treatments can be developed to improve EFL instruction.

Moreover, spoken learner corpora such as YESCOLE can be used in data-driven learning (DDL; Johns 1991), syllabus and material design, and language testing. On the one hand, teachers and students can directly use such corpora via computer software to search for examples of learner language use. Although many teachers may not be willing to integrate corpus directly in their lessons due to time limitations, lack of corpus knowledge and technical constraints such as the absence of computer facilities (Farr 2008; Hedayati and Marandi 2014; Ebrahimi and Faghih 2017), corpus-driven data can become quite beneficial to work on. On the other hand, corpus-driven data can be used indirectly to prepare ELT materials, course syllabi, language tests and exercises. In contrast to the direct use of corpus, corpus consultation may be more favorable especially in technology-poor contexts (Hedayati and Marandi 2014). Concordance lines extracted through a corpus tool can be used in activity preparation. For example, error-tagged corpora such as YESCOLE can be used to create error-correction activities in the classroom or the structures that students have difficulty with can be taught and then tested in the exams. The use of concordance lines obtained via the corpus analysis software, as illustrated in Figure 2, can bring students' attention to common word formation errors (coded as

[*m:affix] in YESCOLE). The concordance lines can be presented either as screenshots or used in an error-correction exercise.

] knowledge and also their knowledge, intelligent [*m:affix] for example, intelligent [*m:affix] knowledge. Se
people can't find a job because of crowded [*m:affix] For example, when somebody is looking for a
[*s:0art] credit card. Also, it's not safety [*m:affix] for me in terms of stealing. And to
I do. All in all, for being freedom [*m:affix] I think online shopping is not a good
marriage, such as love, lifestyle and dangerous [*m:affix] In my own idea, we shouldn't marry we
and another and the last disadvantage is extract [*m:affix] information from people. [13C] [In marriages larg
ing in business life. [*s:0art] First beneficial [*m:affix] is related to economical [*m:affix] and [*s:0art]

Figure 2: Screenshot of sample concordance lines showing some word formation errors in *AntConc*

Accordingly, teachers can give corpus-driven feedback on learners' speaking errors and output, so students will notice and be aware of the most common errors committed. Below is a sample exercise prepared with concordance lines that include word formation errors from YESCOLE.

Error-correction Exercise

The following concordance lines have been taken from YESCOLE, a spoken corpus of Turkish learners of English. Read the lines and correct the words with an asterisk (*).

1. ... because of their tastes and cheap*, also their preparation to an. As a ...
2. ... and, staying away from crowdness* and they can live small ...
3. ... people's thoughts and their speaks*. And also she can do home chores...
4. ... may feel themselves more energic*. And they can enjoy their life their ...
5. ... is related to economical* and second is ...
6. ... in big cities, but it's crowd* and most of the companies...
7. ... trust issue is one of the nature* aspect of unhappy marriage. For example...
8. ... compared to villages. Like sport* centers, parks, stuff like that. And people...
9. ... First of all, it has a deterrent* effect on the society. Many people in...
10. ... beautiful thing to do before die*. Everybody should keep a pet, they should...

Another example may be the use of learners' erroneous spoken performance found in the corpus to prepare awareness-raising activities (e.g. Hobbs 2005). For this purpose, transcripts obtained from recordings of students' spoken English can be used as consciousness-raising activities. Below is a sample speaking task prepared with transcript 5B1T1 obtained from YESCOLE. Similarly, transcripts with audio files can be used to detect mispronunciation. Although pronunciation errors have not been tagged in YESCOLE, such spoken corpora can be used to raise learners' awareness.

Speaking Task

A. Imagine that you are asked to decide whether to continue your education online or in traditional classes next semester. In order to decide, you need to list the characteristics of online classes and traditional classes.

<u>Online Classes</u>	<u>Traditional classes</u>
- online	- face-to-face
- ...	- ...

B. A Turkish learner of English talks about whether online classes are better than traditional classes. Read the transcript of the talk taken from YESCOLE and correct all the errors. How many errors did you find? Compare your answers with those of your classmates.

Hello. Today I'm going to talk about online classes are ... if they are better than conventional classes. I think I... it's not. I don't agree with that idea.

Because, first of all, I cannot make a conversation with my teacher face-to-face. And to explain, when I'm [xxx] in reality not in online, I feel more comfortable and ask whatever I want. Secondly, I cannot ask any questions while I'm not in classes, online classes. Mainly I have some question marks in my head. And I do not ask it to you because we are talking on skype or something other social media applications. What else? Finally, online classes are works with electric like computers and other electronic devices. If electric goes out I cannot keep up with the classes. So maybe I should go.

So online classes are very useless, in my opinion. I cannot make any face-to-face conversation. I cannot ask any question while I'm not in class, online class, and if electric goes out I cannot keep up with my classes. And I can miss some classes.

C. Prepare a similar speech on the same topic. Do you think online classes are better than traditional classes? Justify your answer.

De Moraes (2018) also points out that a spoken learner corpus can be used to teach speaking through creating instructional activities tailored to the needs of a specific learner group. Furthermore, as suggested by Gilquin *et al.* (2007), a specialized spoken corpus can fill the gaps in English for Academic Purposes (EAP) pedagogy to create teaching and testing materials. To illustrate, wordlists can be made through the corpus tool and students' vocabulary profile might be observed; review classes might be organized considering the resistant errors; and quality distractors can be selected from the error-tagged corpus while preparing language tests. In this vein, building and making use of learner corpora provide opportunities for teachers, curriculum/course designers, and test developers to prepare teaching and testing materials using the spoken language of learners.

Lastly, not only the features of spoken English, such as discourse markers, ellipsis, headers and tails but also communication strategies (e.g. compensation speaking strategies) of learners at different levels of proficiency can be investigated with the help

of learner spoken corpora. As can be seen, developmental spoken learner corpora, such as YESCOLE, can contribute to the field of ELT in many respects.

4. CONCLUSION

This paper introduced YESCOLE as a representative, specialized and developmental learner corpus of spoken English. It is a novel source of learner data in the Turkish EFL context which, to the best of our knowledge, is currently the only developmental/longitudinal spoken corpus of English-major EFL learners at different levels of proficiency in this context. In that sense, we believe that YESCOLE includes invaluable spoken data to inspire studies that might reveal significant instructional implications contributing to the field of ELT. The paper describes in detail the steps taken into account to build YESCOLE as summarized below:

- Checking the proficiency level of the learners of English,
- selecting the prompts to elicit oral data,
- planning the data collection timings and preparing the setting for oral recordings,
- transcribing the oral recordings,
- identifying the utterances as units of analysis,
- annotating the corpus, and
- conducting automatic corpus analyses.

These steps can be guiding and inspirational for future researchers who aim to compile a developmental/longitudinal spoken learner corpus as well. We truly hope that the spoken corpus compilation becomes a more common practice in years to come to be able to provide a corpus-based evidence to the language development of EFL learners all over the world.

REFERENCES

- Ahangari, Saeideh and Morteza Abdi. 2011. The effect of pre-task planning on the accuracy and complexity of Iranian EFL learners' oral performance. *Procedia – Social and Behavioral Sciences* 29: 1950–1959.
- Anthony, Laurence. 2017. *AntFileConverter* (version 1.2.1). Tokyo, Japan: Waseda University. <http://www.laurenceanthony.net/software>
- Anthony, Lawrence. 2019. *AntConc* (version 3.5.8). Tokyo, Japan: Waseda University. <http://www.laurenceanthony.net/software>

- Asik, Asuman and Pasa Tevfik Cephe. 2013. Discourse markers and spoken English: Nonnative use in the Turkish EFL setting. *English Language Teaching* 6/12: 144–155.
- Belz, Julie A. and Nina Vyatkina. 2008. The pedagogical mediation of a developmental learner corpus for classroom-based language instruction. *Language Learning & Technology* 12/3: 33–52.
- Boers, Frank, June Eyckmans, Jenny Kappel, Helene Stengers and Murielle Demecheleer. 2006. Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research* 10/3: 245–261.
- Bigi, Brigitte. 2015. SPPAS – Multi-lingual approaches to the automatic annotation of speech. *The Phonetician – International Society of Phonetic Sciences* 111: 54–69.
- Council of Europe. 2005. *Reference Supplement to the Preliminary Version of the Manual for Relating Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. DGIV/EDU/LANG 2005, 13. Strasbourg: Language Policy Division.
- De Cock, Sylvie. 2010. Spoken learner corpora and EFL teaching. In Mari Carmen Campoy-Cubillo, Begona Bellés-Fortuño and M. Luisa Gea-Valor eds. *Corpus-based Approaches to English Language Teaching*. London: Continuum, 123–137.
- De Jong, Nivja H., Margarita P. Steinell, Arjen Florijn, Rob Schoonen and Jan H. Hulstijn. 2013. Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics* 34/5: 893–916.
- De Moraes, Helmaria Febeliana Real. 2018. Use of corpora in teaching speaking. In John I. Lontos ed. *The TESOL Encyclopedia of English Language Teaching*. New York: John Wiley and Sons, 1–6.
- Demirel, Elif Tokdemir and Koray Şahin. 2015. The use of spoken learner corpora to detect problems with lexical accuracy. *HUMANITAS-Uluslararası Sosyal Bilimler Dergisi* 3/5: 73–83.
- Demirel, Elif Tokdemir and Semin Kazazoğlu. 2015. The comparison of collocation use by Turkish and Asian learners of English: The case of TCSE corpus and ICNALE corpus. *Procedia – Social and Behavioral Sciences* 174: 2278–2284.
- Du Bois, John W. 1991. Transcription design principles for spoken discourse research. *Pragmatics* 1/1: 71–106.
- Dulay, Heidi C., Marina K. Burt and Stephen D. Krashen. 1982. *Language Two*. Oxford: Oxford University Press.
- Ebrahimi, Alice and Esmail Faghih. 2017. Integrating corpus linguistics into online language teacher education programs. *ReCALL: The Journal of EUROCALL* 29/1: 120–135.
- Ellis, Rod and Fangyuan Yuan. 2004. The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition* 26/1: 59–84.
- Farr, Fiona. 2008. Evaluating the use of corpus-based instruction in a language teacher education context: Perspectives from the users. *Language Awareness* 17/1: 25–43.
- Gilquin, Gaëtanelle. 2015. From design to collection of learner corpora. In Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier eds. *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 9–34.
- Gilquin, Gaëtanelle, Sylviane Granger and Magali Paquot. 2007. Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes* 6/4: 319–335.

- Granger, Sylviane. 2004. Computer learner corpus research: Current status and future prospects. In Ulla Connor and Thomas A. Upton eds. *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam: Rodopi 123–145.
- Granger, Sylviane, Estelle Dagneaux, Fanny Meunier and Magali Paquot eds. 2009. *International Corpus of Learner English*. Louvain-la-Neuve: Presses universitaires de Louvain
- Hedayati, Hora Fatemeh and S. Susan Marandi. 2014. Iranian EFL teachers' perceptions of the difficulties of implementing CALL. *ReCALL* 26/3: 298–314.
- Hobbs, James. 2005. Interactive lexical phrases in pair interview tasks. In Corony Edwards and Jane Willis eds. *Teachers Exploring Tasks in English Language Teaching*. London: Palgrave Macmillan, 143–156.
- Huang Lan Fen. 2011. *Discourse Markers in Spoken English: A Corpus Study of Native Speakers and Chinese Non-native Speakers*. Birmingham: University of Birmingham dissertation.
- Johns, Tim. 1991. From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In Tim Johns and Philip King eds. *Classroom Concordancing. English Language Research Journal*, 4, Birmingham: University of Birmingham. 1–16.
- Khan, Sarah. 2011. *Strategies and Spoken Production of Three Oral Communication Tasks: A Study of High and Low Proficiency EFL Learners*. Barcelona: Universitat Autònoma de Barcelona dissertation.
- Kilimci, Abdurrahman. 2014. LINDSEI-TR: A new spoken corpus of advanced learners of English. *International Journal of Social Sciences and Education* 4/2: 401–410.
- Kormos, Judit, and Zoltán Dörnyei. 2004. The interaction of linguistic and motivational variables in second language task performance. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 9/2. <https://ojs.tu-journals.ulb.tu-darmstadt.de/index.php/zif/article/view/482/458> (20 May, 2021.)
- MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk* (third edition). Mahwah, NJ: Lawrence Erlbaum Associates.
- Masrom, Umi Kalsom, Nik Aloesnita Nik Mohd Alwi and Nor Shidrah Mat Daud. 2015. The role of task complexity and task motivation in language production. *GEMA Online Journal of Language Studies* 15/2: 33–49.
- McEnery, Tony and Costas Gabrielatos. 2006. English corpus linguistics. In Bas Aarts, April MS McMahon and Lars Hinrichs eds. *The Handbook of English Linguistics*. Oxford: Blackwell, 33–71.
- Meunier, Fanny. 2016. Introduction to the LONGDALE Project. In Erik Castello, Katherine Ackerley and Francesca Coccetta eds. *Studies in Learner Corpus Linguistics. Research and Applications for Foreign Language Teaching and Assessment*. Berlin: Peter Lang, 123–126.
- Ortega, Lourdes. 1999. Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition* 21/1: 109–148.
- Oxford Quick Placement Test* (Version 1). 2001. Oxford University in collaboration with University of Cambridge, Local examinations Syndicate, Oxford: Oxford University Press.
- Rea Rizzo, Camino. 2010. Getting on with corpus compilation: From theory to practice. *ESP World* 9:1–23.
- Selinker, Larry. 1972. Interlanguage. *IRAL-International Review of Applied Linguistics in Language Teaching* 10: 209–232.
- Skehan, Peter. 1998. *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

- Stolcke, Andreas, Elizabeth Shriberg, Rebecca Bates, Mari Ostendorf, Dilek Hakkani, Madelaine Plauche, Gökhan Tur and Yu Lu. 1998. Automatic detection of sentence boundaries and disfluencies based on recognized words. In the *Fifth International Conference on Spoken Language Processing*. Sydney, Australia (November 30-December 4, 1998). https://www.isca-speech.org/archive/archive_papers/icslp_1998/i98_0059.pdf (20 February, 2021.)
- Thewissen, Jennifer. 2013. Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal* 97/1: 77–101.
- Ting, Su-Hie, Mahanita Mahadhir and Siew-Lee Chang. 2010. Grammatical errors in spoken English of university students in oral communication course. *GEMA Online Journal of Language Studies* 10/1: 53–70.
- Tono, Yukio and María Belén Díez-Bedmar. 2014. Focus on learner writing at the beginning and intermediate stages: The ICCI corpus. *International Journal of Corpus Linguistics* 19/2: 163–177.
- Traum, David R. and Peter A. Heeman. 1997. Utterance units in spoken dialogue. In Elisabeth Maier, Marion Mast and Susann LuperFoy eds. *Dialogue Processing in Spoken Language Systems*. Heidelberg: Springer, 125–140.
- Worm, Karsten L. 1998. A model for robust processing of spontaneous speech by integrating viable fragments. In Association for Computational Linguistics eds. *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*, 1403–1407. <https://www.aclweb.org/anthology/P98-2229/> (12 November, 2020.)
- Yaman, Sibel, Li Deng, Dong Yu, Ye-Yi Wang and Alex Acero. 2008. An integrative and discriminative technique for spoken utterance classification. *IEEE Transactions on Audio, Speech, and Language Processing* 16/6: 1207–1214.
- Yıldız, Mustafa. 2016. Contrastive analysis of Turkish and English in Turkish EFL learners' spoken discourse. *International Journal of English Studies* 16/1: 57–74.
- Yuan, Fangyuan and Rod Ellis. 2003. The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied linguistics* 24/1: 1–27.

Corresponding author

Ece Genç-Yöntem
Yeditepe University
Kayışdağı BLVD, 326A
Ataşehir, PO Box 34755
İstanbul, Turkey
e-mail: ece.genc@yeditepe.edu.tr

received: November 2020

accepted: May 2021

published online: 2021

How is information content distributed in RA introductions across disciplines? An entropy-based approach

Wei Xiao – Jin Liu – Li Li
Chongqing University / China

Abstract – Recent years have witnessed a growing interest in research article (RA thereafter) introductions. Most previous studies focused on the macro structures, rhetorical functions and linguistic realizations of RA introductions, but few intended to investigate the information content distribution from the perspective of information theory. The current study conducted an entropy-based study on the distributional patterns of information content in RA introductions and their variations across disciplines (humanities, natural sciences, and social sciences). Three indices, that is, one-, two-, and three-gram entropies, were used to analyze 120 RA introductions (40 introductions from each disciplinary area). The results reveal that, first, in RA introductions, the information content is unevenly distributed, with the information content of Move 1 being the highest, followed in sequence by Move 3 and Move 2; second, the three entropy indices may reflect different linguistic features of RA introductions; and, third, disciplinary variations of information content were found. In Move 1, the RA introductions of natural sciences are more informative than those of the other two disciplines, and in Move 3 the RA introductions of social sciences are more informative as well. This study has implications for genre-based instruction in the pedagogy of academic writing, as well as the broadening of the applications of quantitative corpus linguistic methods into less touched fields.

Keywords – RA introductions; move analysis; CARS model; entropy; information theory; disciplinary differences

1. INTRODUCTION¹

Research articles (RAs) are regarded as a central genre of knowledge production and dissemination, as well as a key medium for the legitimating of claims and disciplines (Berkenkotter and Huckin 1995: 3; Hyland 2000: 175). As a “crafted rhetorical artifact”

¹ This work was presented at the 12th *International Conference on Corpus Linguistics* (CILC 2021). It was supported by the Social Science Foundation of Chongqing under Grant No. 2019QNY51, the Fund of the Interdisciplinary Supervisor Team for Graduate Programs of Chongqing Municipal Education Commission under Grant No. YDSTD1923, the Graduate Research Innovation Program of Chongqing under Grant No. CYS20045 and the Fundamental Research Funds for the Central Universities under Grant No. 2021CDJSKZX07. We would like to extend our sincere gratitude and appreciation to the anonymous reviewers for their comments and suggestions.



and a “manifestation of rhetorical maneuver” (Swales 1990: 155), introductions play a critical role in RAs. The introduction section not only provides the interpretive structure that illustrates how readers may decode a study (Grant and Pollock 2011) but also functions to state the significance of the research field by concisely situating the actual research (Swales 1990: 142). Such salience and identifiability have made introductions the focus of a great number of researchers (e.g. Samraj 2002; Hirano 2009; Loi and Evans 2010; del Saz Rubio 2011).

One line of research in this regard has concentrated on the macro structures of RA introductions. Macroscopically, RA introductions can be regarded as composed of several moves, which work together as functional units in a given text (Connor *et al.* 1995) and can be illustrated by their specific communicative purposes and linguistic devices. Previous researchers have investigated the occurrences, sequences and patterns of moves of RA introductions (Samraj 2002; Fakhri 2004; Kanoksilapatham 2005; Hirano 2009; Loi 2010; Sheldon 2011; Lim 2012; Ahamad and Yusof 2012; Muangsamai 2018; Ye 2019) and have indicated that their macro-organization mainly follows Swales’ (1990, 2004) Create-A-Research-Space (CARS) model, i.e. (1) Establishing a Territory, (2) Establishing a Niche and (3) Occupying the Niche, with some conventional moves appearing more frequently than some others that are elective (Cortes 2013). These studies, from a macro perspective, have unraveled the combination and sequences of rhetorical strategies used in RA introductions.

Another branch of research has dealt with microscopic concerns. Some delved into metadiscourse features, such as interactive and reflexive features (Kashiha and Marandi 2019; Li and Xu 2020), the usage of metadiscourse markers in the rhetorical moves of RA introductions (del Saz Rubio 2011; Khedri and Kritsis 2018) and rhetorical and metadiscoursal variations (Kim and Lim 2013; Validi *et al.* 2016). Del Saz Rubio (2011) analyzed the metadiscoursal features in RA introductions, and found that evidentials, transition markers, and code glosses were the most omnipresent interactive categories. Kim and Lim (2013) examined the use of metadiscourse in RA introductions of educational psychology and found that far more interactive than interactional metadiscourse markers were used. Some others concentrated on lexico-grammatical features, such as lexical bundles (Cortes 2013; Esfandiari and Barbary 2017; Mizumoto *et al.* 2017), phraseological units (Liu and Lu 2020), and linguistic realizations utilized to fulfill a specific rhetorical function in the moves of RA introductions (see Lim 2012; Ädel

2014; Joseph *et al.* 2014; Tankó 2017; Lu *et al.* 2020). For instance, Lim (2012) examined how sophisticated writers in the field of management employ linguistic choices to establish research niches. The findings showed that gap indications in management RA introductions were realized by the employment of such expressions as negative verb phrases and attributive quantifiers demonstrating inadequacy. From the perspective of appraisal and evaluation, Wang and Yang (2015) investigated how the promotion was realized in RA introductions in applied linguistics, and what appeals and linguistic devices could be used to show the significance of the study. Their results show that claiming centrality occupied a prominent role in research promotion and worthiness indication. These explorations are of great significance in that they have deepened our understanding of how these rhetorical and linguistic features of RA introductions function in knowledge construction and communication.

Researchers have also been interested in the variations of RA introductions across disciplines (Holmes 1997; Samraj 2002; Ozturk 2007; Lin and Evans 2012; Kanoksilapatham 2015). Samraj (2002), for example, studied the disciplinary variations in move structures by comparing RA introductions of conservation biology and wildlife behavior. She found that the conservation biology introductions were more likely to use such steps as centrality claims and were more concerned with real-world matter instead of the epistemic world of research than those wildlife behavior introductions. Martín and León Pérez (2014) investigated how researchers in the fields of health sciences and social sciences promote their research in Move 3 (Occupying the Niche) and the corresponding linguistic realizations of each step as well. The results demonstrated that health sciences texts showed a higher degree of rhetorical promotion than those of social sciences. These cross-disciplinary investigations have revealed significant differences in move patterns of RA introductions. They have also promoted the awareness of disciplinary differences in the academic community and shed light on academic writing teaching and training.

Recently there have been a few attempts to analyze texts from the perspective of information theory, a field concerned with the concept and measurement of information (van der Lubbe 1997: 1). Entropy, an index widely used in information theory which allows to measure the information content of a message, has been introduced into language studies. This index is different from some widely used indices in corpus linguistics (e.g. TTR) in that it not only considers the variety of words but also their evenness of distribution (Zhu and Lei 2018). The relevant entropy-based studies have

covered such topics as linguistic complexity (Juola 2008; Lu 2012; Ehret and Szmrecsanyi 2014, 2019), cultural complexity (Juola 2013; Khany and Kafshgar 2016; Zhu and Lei 2018), and the information content of certain linguistic entities (Chen *et al.* 2016). RA introductions, regarded as essentially opening paragraphs for writers to prepare the ground for the research to come (del Saz Rubio 2011), are also worth investigating from this perspective. Considering that different moves in RA introductions are expected to achieve particular communicative functions (Shehzad 2010) and the salience of different moves appears dissimilar, their distribution of information content may differ. However, to the best of our knowledge, no entropy-based study has been conducted on RA introductions and their information distribution patterns, in particular, something that will be addressed in the present study. In addition, because of the shortage of relevant literature, there is not yet a clear understanding of the substantial significance of different entropies in academic texts, especially in RA introductions. For example, Juola (2013) managed to measure the complexity of the American culture by calculating the entropy of different grams in the *Google Books N-gram Corpus* and thought that one-gram entropy is the indicator of lexical complexity, two-gram entropy reveals relationships between two linguistic entities, and three-gram entropy reflects syntactic complexity. In another study, Zhu and Lei (2018) analyzed the speeches and debates from the British parliament and found that different N-grams may indicate distinct linguistic features, which confirmed Juola's (2013) claims. Despite their explorations, they have dealt with limited genres other than RA introductions. Their findings thus may not be conclusive. This study will then employ three indices to measure information content, hoping to uncover the correlates between different entropy indices and RA introduction features. Furthermore, as disciplinary variations have been widely documented in previous studies (e.g. Samraj 2002; Hyland 2000; Ozturk 2007; Kanoksilapatham 2015), there may be different information distribution patterns of RA introductions across disciplines, which awaits exploration. To summarize, the present study aims to answer the following three questions:

- (1) How is information content distributed in RA introduction moves?
- (2) Do different entropy indices reveal different linguistic features of RA introductions?
- (3) Are there any variations in information distributional patterns across disciplines?

2. METHODS

2.1. Corpora

Research articles used in the present study were selected via the Web of Science (WoS) search engine. The collected articles were expected to meet the following criteria: 1) they were from high-ranking journals indicated by the 2019 SCI/SSCI/AHCI indices; 2) they were published in the latest years (2017–2020); 3) they were organized in a typical IMRD (Introduction-Methods-Results-Discussion) structure; 4) they were roughly similar in length (8,000 words); and 5) they were written in English. These standards ensured the representativity, reputation and accessibility of the paper collection (see Nwogu 1997). In the end, 120 research articles were randomly selected: 40 research articles in natural sciences,² social sciences,³ and humanities.⁴ *AntFileConverter* (Anthony 2017), a convenient and free software, was then applied to convert the articles into plain text format. The average length per text of natural sciences, social sciences, and humanities was 8,563, 9,163 and 8,873 words, respectively. One-way ANOVA⁵ and post-hoc tests showed that there was no significant difference at the $p < 0.05$ level in text length across the three disciplines [$F(2, 117) = 0.029, p = 0.97$].

2.2. Information content and entropy

In information theory, the information content of a message can be measured by entropy, an index first introduced by Shannon (1948). Higher entropy indicates more information content. In language studies, the entropy of a given text can be calculated by the following formula, where P_i refers to the probability of the relative frequency of the i th word and N stands for the total number of word types in a given text.

$$H_t = - \sum_{i=1}^N P_i \log_2 P_i$$

² From *Chemical Engineering Journal*, *Computer Communications*, *Microporous and Mesoporous Materials*, *Future Generation Computer Systems* and *Atmospheric Environment*.

³ From *Human Relations*, *International Journal of Research in Marketing*, *International Journal of Information Management*, *World Development* and *Telematics and Informatics*.

⁴ From *Journal of Second Language Writing*, *Journal of Archaeological Science*, *Political Geography*, *Digital Applications in Archaeology and Cultural Heritage* and *Language Learning*.

⁵ One-way ANOVA is a statistical test used to compare means for three or more groups. However, ANOVA can only tell us that there is a significant difference but cannot suggest between which groups the difference exists. The post-hoc test, therefore, should be conducted to make multiple comparisons between every two groups, so as to find out where exactly the difference lies.

As a word can also be considered as a one-gram (Zhu and Lei 2017), the entropy of a word can be regarded as one-gram entropy. Likewise, the formula above can be used to calculate two-gram or three-gram entropy if we replace the set of one-grams by two- or three-grams. Operationally, an n -gram can be identified by extracting every n -word in adjacency within a sentence, as illustrated in example (1).

(1) This is a sentence. And here is another sentence.

In the example above, all the one-grams are *this, is, a, sentence, and, here, is, another, and sentence*. All two-grams are *this is, is a, a sentence, and here, here is, is another, and another sentence*. And all three-grams are *this is a, is a sentence, and here is, here is another, and is another sentence*.

2.3. Move annotation

We adopted Swales' (2004) CARS model as the framework of annotation at the sentence level. Each sentence was analyzed and labeled as Establishing a Territory (Move 1), Establishing a Niche (Move 2), or Occupying the Niche (Move 3). This is illustrated in (2).

(2) The aim of this study is to establish the accuracy of different methods to obtain wave conditions in shallow water for nearshore studies, with a special focus on the wave direction. (de Swart *et al.* 2020: 2).

In example (2), *the aim of this study* refers to a prefacing and preparative expression, informing the readers to focus on the objectives of the study, which serves as a common way for authors to occupy the niche. Therefore, this sentence was labeled as Move 3.

To ensure the coding reliability, two researchers first annotated the sentences independently. The consistency coefficient was as high as 93 percent. The inconsistencies were then left for discussion until a final agreement was reached.

2.4. Data analysis

After the move annotation process, the *R* programming language (version 3.5.0, R Core Team 2018) was used for the extraction of one-, two-, and three-grams, the calculation of entropy values and further data analysis. First, we programmed to extract all the one-,

two-, and three-grams. By removing all the punctuations, we split the texts into words and obtained all the one-grams. For two-grams and three-grams, we split the texts into sentences with punctuation marks as boundaries and extracted all the two-grams and three-grams sentence by sentence. Second, we calculated the probabilities of every one-, two-, and three-gram by dividing the occurrence of a gram by the total grams. Third, we calculated the one-, two-, and three-gram entropies of each move according to the aforementioned formula.

After the calculation of one-, two-, and three-gram entropies of each move, we wrote an *R* script to conduct statistical analysis. First, we used the Shapiro-Wilk test⁶ to check whether the original data were normally distributed. When the data were normally distributed, we applied the ANOVA test, in order to examine whether statistically significant difference exists across moves, grams, or disciplines; we then used the Tukey post-hoc analyses to check precisely where the significant difference appeared. When the data were not normally distributed, we then used the Kruskal-Wallis test to check whether there existed significant between-group difference (indicated by *p* value). The significance level was set at 0.05.

3. RESULTS

The group means of one-gram, two-gram and three-gram entropies across RA introductions moves are presented in Figure 1. The descriptive statistics and the corresponding post-hoc test results are displayed in Table 1. From Figure 1 (A), it can be observed that the information content of the three moves is unevenly distributed. The mean entropy of Move 1 is the highest, followed in sequence by Move 3 and Move 2. The entropy values of two-grams and three-grams, as represented in Figure 1 (B and C), suggest a similar distribution pattern in information content as one-gram entropy unravels. The Kruskal-Wallis test results in Table 1 demonstrate that significant differences appear across moves, and the post-hoc analysis further affirms significant differences across moves ($p < 0.05$), independently of whether we are dealing with one-, two- or three-grams. These results indicate that Move 1 may carry more information and occupy a larger part in RA introductions than the other two moves, whereas Move 2 tends to be less

⁶ The Shapiro-Wilk test is to test whether the data is normally distributed and then to indicate which hypothesis test method is appropriate. If $p > 0.05$, it shows that the data is normally distributed and ANOVA should then be used. If $p < 0.05$, it shows a non-normal distribution and Kruskal-Wallis test should be used.

informative in the introduction part.

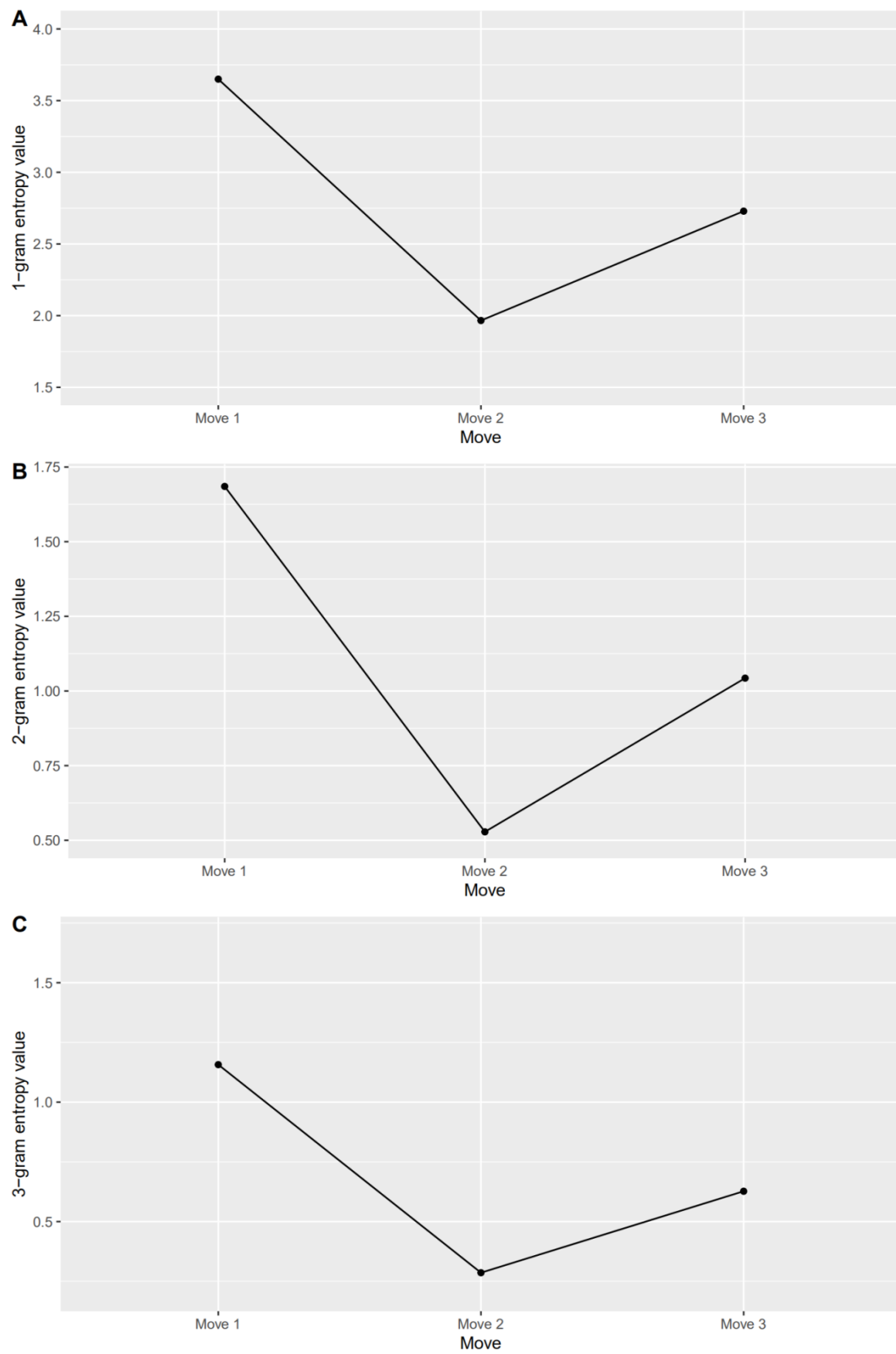


Figure 1: Mean entropy value of one-grams, two-grams, and three-grams across moves

	Mean (SD in parentheses)			X^2	p		Z		
	M1	M2	M3				M1-M2	M2-M3	M1-M3
one-grams	3.650(0.753)	1.966(0.605)	2.729(0.846)	174.27***	0.000	13.191***	-	6.149***	7.042***
two-grams	1.685(0.761)	0.528(0.294)	1.043(0.575)	171.3***	0.000	13.081***	-	6.904***	6.178***
three-grams	1.157(0.671)	0.286(0.191)	0.627(0.406)	176.7***	0.000	13.290***	-	6.889***	6.401***

Table 1: Kruskal-Wallis test results of entropy values across moves⁷

The fact that different indices yield similar information distribution patterns can also be seen in Figure 2 and is indicated by the r values of correlation analyses in Table 2, where there are correlations among the values of one-, two-, and three-gram entropies ($rs > 0.8$, $ps < 0.05$), confirming the consistency of the three indices. Despite their similarities, the indices show some differences in terms of values, as indicated by the p values of pairwise t-tests in Table 3 ($ps < 0.05$). These results suggest that entropies of different grams, though highly correlated, may reveal different linguistic features of RA introductions.

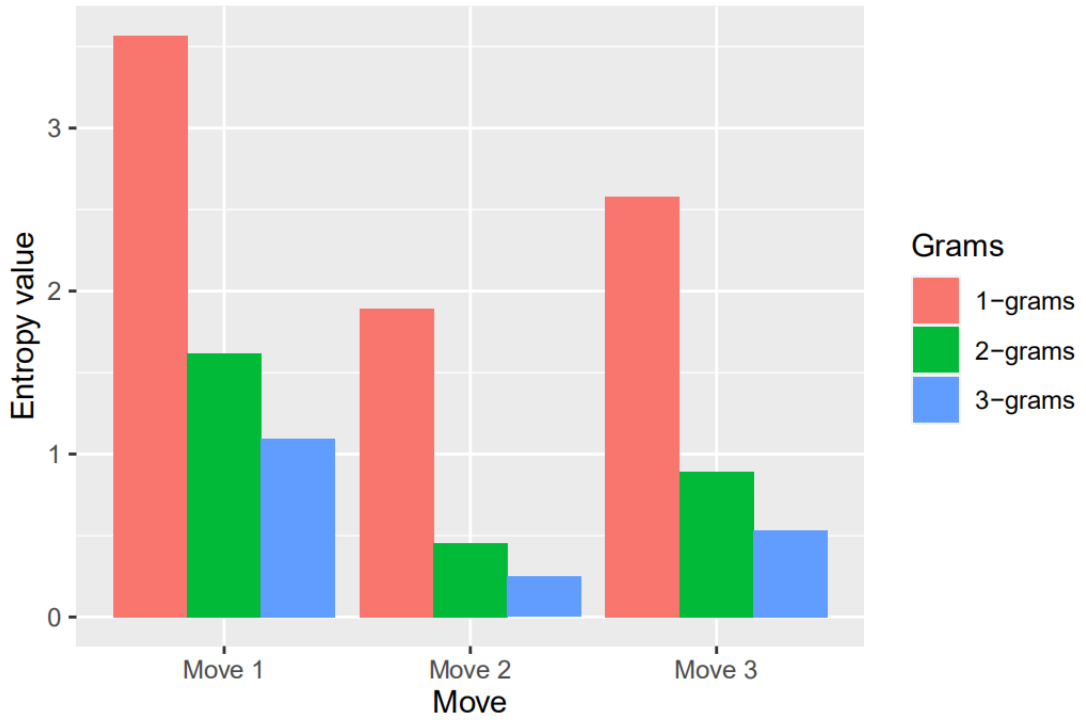


Figure 2: Comparisons among one-, two-, and three-gram entropies

⁷ In this table, three asterisks (***) mean $p < 0.001$.

Move	one-grams - two-grams	two-grams - three-grams	one-grams - three-grams
M 1	0.915***	0.975***	0.852***
M 2	0.897***	0.962***	0.841***
M 3	0.944***	0.964***	0.874***

Table 2: Pearson correlation coefficients between one-, two-, and three-gram entropy values in each move

Move	one-grams - two-grams	two-grams - three-grams	one-grams - three-grams
M 1	1.964***	0.529***	2.493***
M 2	1.437***	0.242***	1.679***
M 3	1.686***	0.416***	2.102***

Table 3: Pairwise t-tests between one-, two-, and three-gram entropy values in each move

Figure 3 represents the information distribution patterns of RA introductions across disciplinary areas (natural sciences, social sciences, and humanities). It can be observed that, in general, RA introductions across the three areas share a similar distribution pattern in information content. The mean of one-gram entropy of Move 1, given any discipline, is much higher than that of Move 3, followed by an even lower entropy of Move 2. The general informative distribution patterns of two-grams and three-grams, to a large extent, resemble that of one-grams. These results indicate that RA introductions from different domains share similarities in information distribution patterns.

With regard to cross-disciplinary variations, the information content of Move 1 of natural sciences is the highest, as indicated by the higher entropies of Move 1. The entropies of Move 1 of humanities and social sciences seem to be much lower. However, the p values of the ANOVA test in Table 4 suggest that there is no interdisciplinary variation in this move ($ps > 0.05$). Despite higher entropy values of social sciences in Move 2, no significant cross-disciplinary variations were found in this move, neither ($ps > 0.05$). In fact, significant differences across disciplines only occur in Move 3, as indicated by the p values of the ANOVA tests (see Table 4). Post-hoc tests further indicate significant differences among social sciences and the other two disciplines, where the mean entropy values of social sciences are considerably higher than that of natural sciences and humanities. Interestingly, there is no significant difference between natural sciences and humanities ($ps > 0.05$).

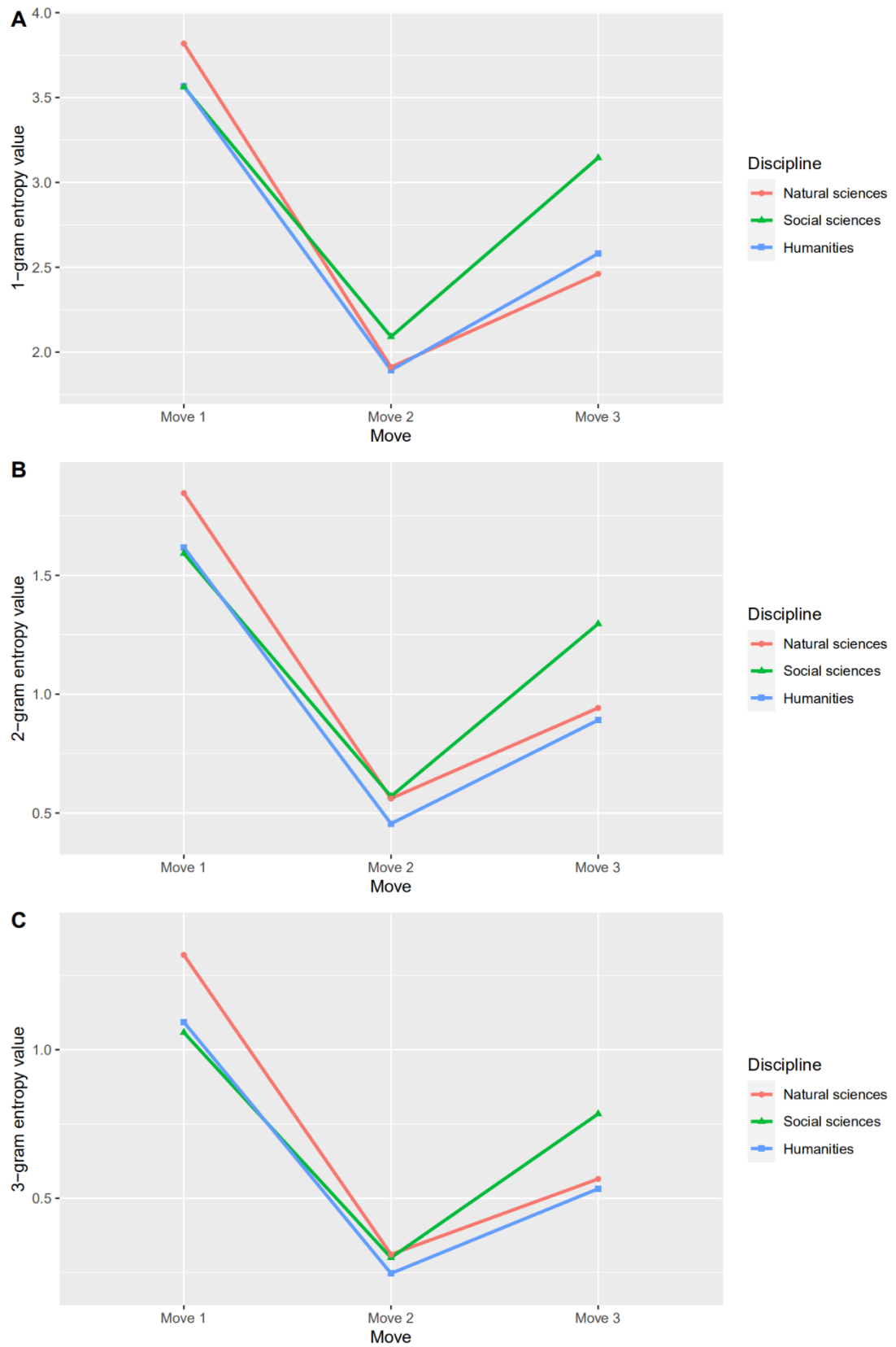


Figure 3: Entropy values of one-grams, two-grams, and three-grams across disciplines

		Mean (SD in parenthesis)			Mean difference				
		NS	SS	HM	<i>F</i>	<i>p</i>	NS-SS	NS-HM	SS-HM
one-grams	M1	3.818 (0.664)	3.563 (0.690)	3.568 (0.878)	1.507	0.226	0.255	0.249	-0.005
	M2	1.912(0.595)	2.091 (0.638)	1.894 (0.577)	1.294	0.278	-0.178	0.018	0.197
	M3	2.461 (0.655)	3.145 (0.880)	2.581 (0.841)	8.359***	0.000	-0.683***	-0.120	0.564***
two-grams	M1	1.846 (0.810)	1.592 (0.726)	1.618 (0.740)	1.354	0.262	0.254	0.228	-0.027
	M2	0.561 (0.293)	0.570 (0.319)	0.455 (0.261)	1.926	0.150	-0.009	0.106	0.115
	M3	0.942 (0.444)	1.296 (0.663)	0.891 (0.523)	6.407**	0.002	-0.354*	0.050	0.405**
three-grams	M1	1.319 (0.751)	1.058 (0.610)	1.093 (0.630)	1.808	0.167	0.261	0.226	-0.036
	M2	0.311 (0.190)	0.300 (0.202)	0.247 (0.178)	1.283	0.281	0.011	0.064	0.052
	M3	0.565 (0.302)	0.784 (0.500)	0.532 (0.352)	4.858**	0.009	-0.219*	0.033	0.252*

Table 4: Cross-disciplinary variations of entropy values⁸

4. DISCUSSION

The current study aimed to explore the information features of RA introductions, focusing on the information content distribution across moves, the linguistics features reflected by different entropy indices, and the cross-disciplinary variations. The moves of RA introductions were identified according to the CARS model, and their one-, two-, and three-gram entropies were calculated and compared. Our findings show that the information content is unevenly distributed across moves, where Move 1 takes up the largest proportion, followed in sequence by Move 3 and Move 2. Besides, the information measured by the three indices represents a similar pattern, although both similarities and variations were found across disciplines.

4.1. The informative distributional pattern of RA introduction moves

The information content in RA introductions was found to be unevenly distributed. The unevenness could be ascribed to the distinctive rhetorical functions of moves and their variations in salience in an RA introduction.

Move 1 (Establishing a Territory) was found to be the most informative one.

⁸ In this table, one asterisk (*) means $p < 0.05$, two asterisks (**) imply $p < 0.01$, and three asterisks (***) are used to indicate $p < 0.001$; NS stands for natural sciences, SS for social sciences, and HM for humanities.

According to Swales' (1990) CARS model, RA introductions often begin with a move that aims to assure readers that the general topic being discussed is worth investigating and that the field of the study is well established through an exhaustive review of the previous studies in the area. As Lim (2012) suggested, Move 1 constitutes a fundamental rhetorical move in RAs in various domains, and researchers are required to acquire sufficient background knowledge to meet the expectations of the academic community. Saricaoglu *et al.* (2021) also highlighted that reference to previous literature is a defining feature of typically all research writings. The literature review or background information thus plays a critical role and occupies a large amount of content in RA introductions. Therefore, it is rather understandable that Move 1 takes up the highest information proportion in RA introductions.

Move 2 (Establishing a Niche), serving as the hinge that connects Move 1 with Move 3 (Swales and Feak 2012: 348), seems to contain much less information, which can also be accounted for by the rhetorical function this specific move bears on. In Move 2, writers are supposed to utilize evaluative resources, mostly negative evaluations or even criticisms, to assess previous literature to create a research niche (Lim 2012). According to Ahmad (1997), Move 2 is used to identify the validation of the piece of research through the description of perceived limitations of the research field or summarizing the work of others. This primary function of Move 2 is often realized through short statements (Xie 2017). As Lindeberg (2004: 139) suggests, Move 2, functioning as a 'mini-critique', shows a preference for conciseness and often entails no more than one sentence, thus occupying a relatively short length in the introduction section. Swales and Feak (2004) also recommend that gap indications in Move 2 should be simple and fairly short due to their easy-to-follow and straightforward manner. Besides, although the way of revealing these limitations is variable, either counter-claiming, indicating a gap in previous studies, question-raising, or continuing a tradition, the phrases and expressions are relatively rigorous and tend to have more syntactic regularities. Shehzad (2008), for instance, pointed out that the gap statements were often realized by contrastive markers (e.g. *nevertheless, however*, etc.), quantifiers (e.g. *relatively few*), and negative statements (e.g. *none, no study*, etc.). This precise and rigorous nature may lead to the comparatively less informative proportion of Move 2.

Another possibility of the lower content of information in Move 2 could be related to socio-cultural factors. Due to the significance of face-saving in academic communities,

researchers may find it difficult and inappropriate to identify previous studies and point out the possible shortcomings and limitations that previous works might have (Taylor and Chen 1991; Kanoksilapatham 2005; Hamp-Lyons and Heasley 2006: 45). Taylor and Chen (1991), for example, found that the authors of Chinese papers were less likely to evaluate previous studies and provide less extensive discussions of other scholars' works owing to the unacceptability of argumentations and confrontations in the Chinese socio-cultural context. The reasons mentioned above may also contribute to the less information contribution of Move 2.

As for Move 3 (Occupying the Niche), the primary function is to turn the niche established in Move 2 into the research space that identifies the present research (Swales 1990). Essential components of this move include descriptions of the research purposes and the current work to be carried out, announcements of principal findings, and sometimes, presentations of the organization of the present paper. As Martín and León Pérez (2014) state, it is in Move 3 that scholars essentially present their research by outlining the research purpose. Move 3 plays a vital role in convincing peers of the relevance and validation of the study to be conducted, in which various promotional strategies are used to highlight the value of their work. Writers concentrate on the use of various rhetorical strategies to promote their 'selling point' (Hyland 2000; Shehzad 2010) through anticipating the principal findings or highlighting the significance, newness, and contributions that their work makes to the field. Shehzad (2010) also suggests that the introductions in computer sciences are result-oriented, where scholars would elaborate explanations in the description of results and highlight the writers' contributions in various ways. Scholars in this field conduct elaborate explanations in the description of results and in various ways highlight the writers' contributions to the field.

Since different moves play different roles in the organization of an RA introduction and the focus of differing moves is not identical, it is no wonder that variations in information distribution across moves exist.

4.2. The informative variations across different grams

The information distribution patterns revealed by different indices are of great similarity. This finding is in accordance with that of Zhu and Lei (2018), in which similar patterns were also found among the one-, two-, and three-gram entropies, thus corroborating the

usefulness of entropy as a measurement of information content. Despite the similarity, our results are different from those of Zhu and Lei (2018) in that we found larger one-gram entropy values than two- and three-gram ones, while they found the opposite. The possible reasons are twofold. First, the data employed in our study are academic articles while Zhu and Lei (2018) collected materials of spoken texts (speeches of the British Parliament). Thus, the divergence may be explained across genres. Second, this difference could be ascribed to the authentic and scientific nature of RAs, which are featured with prescriptive organization patterns (Ren and Li 2011; Taş 2008). There are a variety of recommended sentence patterns or conventional phrases for writers' reference (Cross and Oppenheim 2006). In order to be accepted by the academic communities, scholars have to conform to these academic conventions and increasingly show a preference for a restricted repertoire of identified moves (Holmes 1997). Hence, fixed patterns seem to lead to a lower syntactic complexity. As Juola (2013) suggested, one-gram entropy reflects lexical complexity, whereas two-gram entropy reveals relationships between two entities, and three-gram entropy is more concerned with syntactic complexity. Thus, it is explainable that the entropy value of two- and three-grams are considerably lower than that of one-grams.

4.3. The information distribution patterns of RA introduction moves across disciplines

With regard to disciplinary variations, significant differences can be found in Move 3, where the information content of social sciences was considerably higher than those of the other two disciplines. This finding could be accounted for by the discursive nature of social sciences. As Kuteeva and Airey (2014) state, unlike natural sciences (also regarded as hard sciences) that stress the quantitative and experimental nature of materials, social sciences (typically considered as soft sciences) emphasize the qualitative and interpretive disposition. In social sciences, knowledge is normally regarded as a process of making interpretations of a certain issue or some social phenomenon. The way of interpretation is very important in the process of persuasion of readers. Writers of social sciences may have diverse personal writing styles, which can be reflected through the employment of new words or innovative expressions, thus contributing to the informativeness of the text. Our findings are consistent with some previous studies. Lu *et al.* (2020), for example, investigated the rhetorical functions of syntactically complex sentences in RA introductions of social sciences and found that writers in social sciences tend to elaborate

on announcing the piece of research by the use of a variety of structures at the phrasal and clausal levels, giving rise to greater syntactic complexity. The occurrence of more new word types in Move 3 could be an indication of the discursive nature of social sciences, which is captured by the higher entropy values.

It should also be noted that the information content of Move 1 in natural sciences is higher than that of social sciences and humanities, although no significant difference appears. The emphasis on Move 1 can be explained by the accumulative and iterative nature of natural sciences. In natural sciences, the processes of knowledge construction rely heavily on solid foundations accumulated by experimental support or empirical data (Kuteeva and Airey 2014). Compared with those in social sciences and humanities, works in natural sciences are typically featured with shared paradigms, in which references to previous literature are of great importance. Our findings are in accordance with some previous studies. Samraj (2008), for instance, explored the introductions from three disciplines (linguistics, philosophy, and biology) and found that biology students tend to establish stronger links to previous works. Another possibility could be related to the richness of content words in natural sciences. Due to the typical feature of ready-made paradigms in natural sciences, researchers have to resort to an essential number of prefabricated concepts and technical terms, hoping to establish a common background with peers in the same field, and to be accepted by the academic community (Xiao and Sun 2020). In addition, as science and technology are rapidly developing, a flood of new terms and concepts keeps emerging in this field, thus contributing to the complexity in establishing a territory. These reasons may give rise to a higher information proportion in the Move 1 of natural sciences.

5. CONCLUSION

The present study investigated the information contents of RA introductions and the variations across disciplines. An entropy-based approach was employed and three quantitative indices, that is, one-, two-, and three-gram entropies, were adopted. It was found that the information content is unevenly distributed in introductions, where Move 1 tends to be more informative than the other two moves. It was also found that the one-gram entropy values were higher than the other two indices. Furthermore, disciplinary variations were also attested. In Move 1, the RA introductions of natural sciences are more informative than those of the other two disciplines, and in Move 3 the RA

introductions of social sciences are more informative. These differences may be explained by the rhetorical and linguistic features of individual moves, the different aspects possibly reflected by different indices and the very nature of distinctive disciplines.

This study is a preliminary attempt to explore RA introductions from the perspective of information theory. It demonstrates the promising prospects of using quantitative linguistics methods in the study of RA introductions and many other genres, where traditional qualitative methods and basic statistics still prevail. The findings suggest how information content is supposed to be distributed and arranged in RA introductions, which is difficult to find with mere qualitative methods. This study also sheds light on the pedagogy of academic writing, in that it can help students to be aware of the structure of introductions and the prominence of each move, as well as the potential conventions and paradigms of their own disciplines. With this awareness in mind, students would be better equipped in the organization of introductions, the allocation of information and the promotion of research, which will eventually lead to successful academic writing and reputation winning.

There are also several limitations in this study. First, due to the laborious manual coding processes, we only analyzed 120 introductions. Future studies may enlarge the corpus size to ensure the stableness of results. Second, different disciplines and even sub-disciplines may have their own features, which threatens the homogeneity of constitution and challenges the representativeness and generalizability of sampling and corpus building. Future studies may use data from a wider range of disciplines to cross validate our results. Third, although we managed to reveal the significant differences of entropies of different grams in RA introductions, more fine-grained measurements, that is, the four-, five-, six-gram entropies were not explored, which awaits further research. Fourth, we only investigated the introduction part of RAs. Future studies may extend the focus to other parts, such as abstracts, to unravel more features of information distribution in research articles.

REFERENCES

- Ädel, Annelie. 2014. Selecting quantitative data for qualitative analysis: A case study connecting a lexicogrammatical pattern to rhetorical moves. *Journal of English for Academic Purposes* 16: 68–80.

- Ahamad, Mohamed I. and Amira M. Yusof. 2012. A genre analysis of Islamic academic research article introductions. *Procedia - Social and Behavioral Sciences* 66: 157–168.
- Ahmad, Ummul. 1997. Research article introductions in Malay: Rhetoric in an emerging research community. In Anna Duszak ed. *Culture and Styles in Academic Discourse*. Berlín: Mouton de Gruyter, 273–303.
- Anthony, Laurence. 2017. *AntFileConverter* (version 1.2.1). Tokyo, Japan: Waseda University. <http://www.laurenceanthony.net/> (01 May, 2020.)
- Berkenkotter, Carol and Thomas N. Huckin. 1995. *Genre knowledge in disciplinary communication: Cognition/Culture/Power*. New Jersey: Lawrence Erlbaum Associates.
- Chen, Ruina, Haitao Liu and Gabriel Altmann. 2016. Entropy in different text types. *Digital Scholarship in the Humanities* 32/3: 528–542.
- Connor, Ulla, Kenneth Davis and Teun de Rycker. 1995. Correctness and clarity in applying for overseas jobs: A cross-cultural analysis of US and Flemish applications. *Text & Talk* 15/4: 457–475.
- Cortes, Viviana. 2013. The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes* 12/1: 33–43.
- Cross, Cate and Charles Oppenheim. 2006. A genre analysis of scientific abstracts. *Journal of Documentation* 62: 428–446.
- De Swart, Rinse, Francesca Ribas, Daniel Calvete, Aart Kroon, and Alejandro Orfila. 2020. Optimal estimations of directional wave conditions for nearshore field studies. *Continental Shelf Research* 196: 104071.
- Del Saz Rubio, M. Milagros. 2011. A pragmatic approach to the macro-structure and metadiscoursal features of research article introductions in the field of Agricultural Sciences. *English for Specific Purposes* 30/4: 258–271.
- Ehret, Katharina and Benedikt Szmrecsanyi. 2016. An information-theoretic approach to assess linguistic complexity. In Raffaella Baechler and Guido Seiler eds. *Complexity, Isolation, and Variation*. Berlin: Boston De Gruyter, 71–94.
- Ehret, Katharina and Benedikt Szmrecsanyi. 2019. Compressing learner language: An information-theoretic measure of complexity in SLA production data. *Second Language Research* 35/1: 23–45.
- Esfandiari, Rajab and Fatima Barbary. 2017. A contrastive corpus-driven study of lexical bundles between English writers and Persian writers in psychology research articles. *Journal of English for Academic Purposes* 29: 21–42.
- Fakhri, Ahmed. 2004. Rhetorical properties of Arabic research article introductions. *Journal of Pragmatics* 36/6: 1119–1138.
- Grant, Adam M. and Timothy G. Pollock. 2011. Publishing in AMJ—part 3: Setting the hook. *Academy of Management Journal* 54/5: 873–879.
- Hamp-Lyons Liz and Ben Heasley. 2006. *Study Writing: A Course in Writing Skills for Academic Purposes*. Cambridge: Cambridge University Press.
- Hirano, Eliana. 2009. Research article introductions in English for specific purposes: A comparison between Brazilian Portuguese and English. *English for Specific Purposes* 28: 240–250.
- Holmes, Richard. 1997. Genre analysis, and the social sciences: An investigation of the structure of research article discussions sections in three disciplines. *English for Specific Purposes* 16: 321–337.
- Hyland, Ken. 2000. *Disciplinary Discourse: Social Interactions in Academic Writing*. London: Longman.

- Joseph, Renu, Jason Miin-Hwa Lim and Nor Arifah Mohd. 2014. Communicative moves in forestry research introductions: Implications for the design of learning materials. *Procedia - Social and Behavioral Sciences* 134: 53–69.
- Juola, Patrick. 2008. *Assessing Linguistic Complexity*. Amsterdam: John Benjamins.
- Juola, Patrick. 2013. Using the Google N-Gram corpus to measure cultural complexity. *Literary and Linguistic Computing* 28/4: 668–675.
- Kanoksilapatham, Budsaba. 2005. Rhetorical structure of biochemistry research articles. *English for Specific Purposes* 24/3: 269–292.
- Kanoksilapatham, Budsaba. 2015. Distinguishing textual features characterizing structural variation in research articles across three engineering sub-discipline corpora. *English for Specific Purposes* 37: 74–86.
- Kashiha, Hadi and Susan S. Marandi. 2019. Rhetoric-specific features of interactive metadiscourse in introduction moves: A case of discipline awareness. *Southern African Linguistics and Applied Language Studies* 37/1: 1–14.
- Khany, Reza, and Neda Babanezhad Kafshgar. 2016. Analyzing texts through their linguistic properties: A cross-disciplinary study. *Journal of Quantitative Linguistics* 23/1: 278–294.
- Khedri, Mohsen and Konstantinos Kritsis. 2018. Metadiscourse in applied linguistics and chemistry research article introductions. *Journal of Research in Applied Linguistics* 9/2: 47–73.
- Kim, Loi Check and Jason Miin-Hwa. 2013. Metadiscourse in English and Chinese research article introductions. *Discourse Studies* 15/2: 129–146.
- Kuteeva, Maria and John Airey. 2014. Disciplinary differences in the use of English in higher education: Reflections on recent policy developments. *Higher Education* 67/5: 533–549.
- Li, Zhijun and Jinfen Xu. 2020. Reflexive metadiscourse in Chinese and English sociology research article introductions and discussions. *Journal of Pragmatics* 159: 47–59.
- Lim, Jason Miin-Hwa. 2012. How do writers establish research niches? A genre-based investigation into management researchers' rhetorical steps and linguistic mechanisms. *Journal of English for Academic Purposes* 11/3: 229–245.
- Lin, Ling and Stephen Evans. 2012. Structural patterns in empirical research articles: A cross-disciplinary study. *English for Specific Purposes* 31/3: 150–160.
- Lindeberg, Ann C. 2004. *Promotion and Politeness: Conflicting Scholarly Rhetoric in Three Disciplines*. Pargas, Finland: Åbo Akademi University Press.
- Loi, Chek Kim and Moyra Sweetnam Evans. 2010. Cultural differences in the organization of research article introductions from the field of educational psychology: English and Chinese. *Journal of Pragmatics* 42/10: 2814–2825.
- Lu, Xiaofei. 2012. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly* 45/1: 36–62.
- Lu, Xiaofei, J. Elliott Casal and Yingying Liu. 2020. The rhetorical functions of syntactically complex sentences in social science research article introductions. *Journal of English for Academic Purposes* 44: Article 100832.
- Martín, Pedro and Isabel K. León Pérez. 2014. Convincing peers of the value of one's research: A genre analysis of rhetorical promotion in academic texts. *English for Specific Purposes* 34/1: 1–13.
- Mizumoto, Atsushi, Hamatani Sawako and Imao Yasuhiro. 2017. Applying the bundle-move connection approach to the development of an online writing support tool for research articles. *Language Learning* 67/4: 885–921.

- Muangsamai, Pornsiri. 2018. Analysis of moves, rhetorical patterns and linguistic features in New Scientist articles. *Kasetsart Journal of Social Sciences* 39/2: 236–243.
- Nwogu, Kevin Ngozi. 1997. The medical research paper: Structure and functions. *English for Specific Purposes* 16/2: 119–138.
- Ozturk, Ismet. 2007. The textual organization of research article introductions in applied linguistics: Variability within a single discipline. *English for Specific Purposes* 26: 25–38.
- R Core Team. 2018. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/> (01 May, 2020.)
- Ren, Hongwei, and Yuying Li. 2011. A comparison study on the rhetorical moves of abstracts in published research articles and master's foreign-language theses. *English Language Teaching* 4/1: 162–166.
- Samraj, Betty. 2002. Introductions in research articles: Variations across disciplines. *English for Specific Purposes* 21/1: 1–17.
- Samraj, Betty. 2008. A discourse analysis of master's theses across disciplines with a focus on introductions. *Journal of English for Academic Purposes* 7/1: 55–67.
- Saricaoglu, Aysel, Zeynep Bilki, and Lia Plakans. 2021. Syntactic complexity in learner-generated research paper introductions: Rhetorical functions and level of move/step realization. *Journal of English for Academic Purposes* 53: Article 101037.
- Shannon, Claude E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27/3: 379–423.
- Shehzad, Wasima. 2008. Move two: Establishing a niche. *Iberica* 15/1: 25–50.
- Shehzad, Wasima. 2010. Announcement of the principal findings and value addition in computer science research papers. *Iberica* 19/1: 97–118.
- Sheldon, Elena. 2011. Rhetorical differences in RA introductions written by English L1 and L2 and Castilian Spanish L1 writers. *Journal of English for Academic Purposes* 10/4: 238–251.
- Swales, John M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.
- Swales, John M. 2004. *Research Genres: Explorations and Applications*. Cambridge: Cambridge University Press.
- Swales, John M. and Christine B. Feak. 2004. *Academic Writing for Graduate Students: Essential Tasks and Skills*. Ann Arbor: University of Michigan Press.
- Swales, John M. and Christine B. Feak. 2012. *Academic Writing for Graduate Students*. Ann Arbor: University of Michigan Press.
- Tankó, Gyula. 2017. Literary research article abstracts: An analysis of rhetorical moves and their linguistic realizations. *Journal of English for Academic Purposes* 27: 42–55.
- Taş, Elvan Eda. I. 2008. *A Corpus-based Analysis of Genre-specific Discourse of Research: The PhD Thesis and the Research Article in ELT*. Ankara, Turkey: Middle East Technical University dissertation.
- Taylor, Gordon and Chen Tingguang. 1991. Linguistic, cultural, and subcultural issues in contrastive discourse analysis: Anglo-American and Chinese scientific texts. *Applied Linguistics* 12/3: 319–336.
- Validi, Mahmood, Alireza Jalilifar, Zohreh G. Shooshtari and Abdolmajid Hayati. 2016. Medical research article introductions in Persian and English contexts: Rhetorical and metadiscoursal differences. *Journal of Research in Applied Linguistics* 7/2: 73–98.

- Van der Lubbe, Jan C. A. 1997. *Information Theory*. Cambridge: Cambridge University Press.
- Wang, Weihong and Chengsong Yang. 2015. Claiming centrality as promotion in applied linguistics research article introductions. *Journal of English for Academic Purposes* 20: 162–175.
- Xiao, Wei, and Shuyi Sun. 2020. Dynamic lexical features of PhD theses across disciplines: A text mining approach. *Journal of Quantitative Linguistics* 27/2: 114–133.
- Xie, Jianping. 2017. Evaluation in moves: An integrated analysis of Chinese MA thesis literature reviews. *English Language Teaching* 10/3: 1–20.
- Ye, Yunping. 2019. Macrostructures and rhetorical moves in energy engineering research articles written by Chinese expert writers. *Journal of English for Academic Purposes* 38: 48–61.
- Zhu, Haoran and Lei Lei. 2017. British cultural complexity: An entropy-based approach. *Journal of Quantitative Linguistics* 25/2: 190–205.
- Zhu, Haoran and Lei Lei. 2018. Is modern English becoming less inflectionally diversified? Evidence from entropy-based algorithm. *Lingua* 216: 10–27.

Corresponding author

Wei Xiao
Chongqing University
School of Foreign Languages and Cultures
No.55 Daxuecheng South Rd.
401331. Chongqing
China
e-mail: xiaoweiyx@126.com

received: August 2021
accepted: November 2021

Libya, the media and the language of violence: A Corpus-Assisted Discourse Analysis

Safa Attia
Autonomous University of Madrid / Spain

Abstract – The Arab revolution euphoria of 2011 was covered around the clock by different media sites, engaging millions of followers around the world, and eventually turning into discontent in some affected countries. This study examines the outcomes of the Libyan uprising (2011–2015), specifically the topics of civil-war and terrorism, through the lenses of the Arab written media in Arabic (*Al Jazeera* and *Al Arabiya*), the Arab written media in English (*Al Jazeera* and *Al Arabiya*), and the Western written media in English (BBC and CNN). Through Corpus-Assisted Discourse Analysis (CADS), integrating discursive news values analysis (DNVA), this study highlights the ideological representations of these media, and examines their similarities and differences in terms of frequency distribution and story content. The findings indicate that the media coverage of the outcomes of the Libyan Revolution, when reporting on the topics of war and terrorism, follow similar directions in the story content and the frequency distribution, with some differences in the latter between the analysed media sites. Also, the collocations, concordances, and DNVA results, especially NEGATIVITY, IMPACT and ELITENESS, prove the emphasis of the media on violent language, making terrorism appear the norm, and thus manipulating the audience and affecting their understanding of the news.

Keywords – corpus linguistics; CADS; DNVA; the Libyan revolution; ideology; media violence

1. INTRODUCTION

In 2011, a revolutionary wave swept the Arab regions, calling for the toppling of the old regimes and asking for dignity, employment and freedom, which came to be called ‘the Arab Spring’. Each of the participating countries took a different path, as Haider (2016) lists in his study. In the case of Libya, we find that it endured the fall of the ruling regime and the civil war, which makes it a unique case in comparison with the other countries (Smits *et al.* 2013). These outcomes were reported in all the media, but from different perspectives. This investigation is based on an analysis of the language of the media, specifically *Al Jazeera* and *Al Arabiya* Arabic and English, CNN and BBC in English, as well as the ideological orientations that served their agendas. We used Corpus-Assisted



Critical Discourse Analysis (CADS) in this study (Stubbs 1996; Partington 2006; Taylor 2013), which is a combination of Corpus Linguistics (CL) tools (i.e. frequency, keywords, collocation and concordance) and Critical Discourse Analysis (CDA). We integrated the discursive news values framework (DNVA) with these corpus techniques to enhance the objectivity of the qualitative analysis. In this study we seek to answer four research questions:

1. From a quantitative angle, is there any significant difference in terms of distributions or lexical frequencies between the three categories of news articles?
2. From a qualitative angle, do the salient words reveal the same story contents within the three analysed media categories?
3. What do the collocations, concordances and DNVA reveal about the agendas and ideological orientations of the media organisations?
4. What news values are discursively constructed in English and Arabic?

The originality of this study lies in addressing the war and terrorism issues in Libya during the post-revolution period, not only in three different types of media but also in two languages. Besides, this study is the first to research the Arabic language with a combination of CL, CDA and DNVA approach.

2. LITERATURE REVIEW

2.1. Empirical background

2.1.1. The language of the media

A plethora of empirical research studies has investigated the different ideological orientations of the written media, which are of great significance in relation to the political discourse and play a fundamental role in changing facts and beliefs. Thus, previous studies have shown how each type of media tries to shape public opinion and escalate or de-escalate conflicts. For example, Romero-Trillo (2011) compared the English-language Indian and Pakistani press during their military confrontation in December 2008, using a CDA and a CL analysis. The results revealed the role of both media in igniting the conflict.

Several studies have proved that the media, especially the written media, tend to implement a violent discourse in their news. In fact, Romero-Trillo and Attia (2016) prove that the Arab media, represented by *Al Jazeera* (Arabic and English) and *Al Arabiya*

(Arabic and English), tended to adopt a violent discourse in reporting the events of the Tunisian revolution (2011–2015). The researchers used CDA to examine the outcomes of the Tunisian revolution through the eyes of the Arab and Western Media. The latter, represented by the BBC and CNN, proved, to a certain extent, to be more objective. Nevertheless, Haigh and Bruce (2017), who investigated the visual and story frames employed by *Al Jazeera* English and CNN during the Egyptian uprising (2011), found that both networks tended to use conflict and violence discourse, which tallies with the results obtained by Liebes and First (2004), who conducted a parallel study and revealed the excessive use of conflict images by media.

The present research offers an analysis of the written media coverage of the Arab Spring protests. It is worth mentioning the example of Syria's destructive war, starting in 2011, which has been in sharp focus from all categories of media as the great human disaster of the twenty-first century (Doucet 2018). Additionally, Yılmaz and Sinanoğlu (2014) explored the ideology of two Turkish newspapers, *Sabah* and *Cumhuriyet*, in covering Syria's war and their role in shaping public perception. The findings revealed that *Cumhuriyet* has more intense coverage on impactful, unexpected and brutal news than *Sabah*. However, both newspapers were found to be ideologically biased, which in turn affects the readers' grasp of the news.

Likewise, Haider (2016) compared two Gulf countries' newspapers, namely *Asharq Al-Awsat* and *Al-Khaleej* from 2009 to 2013, in the representation of Muammar Gaddafi before and after the Libyan revolution (2011). The findings display a positive portrayal of Gaddafi before the uprising and an extremely negative image in the post-uprising era, highlighting his cruel behaviour against his people and his involvement in terrorist activities. This, in turn, proves these newspapers' alignment and their effect on the public ideology.

In the same vein, Hamdy and Gomaa (2012) examined Egyptian semi-official newspapers, independent newspapers and grassroots media postings (*Twitter*, *Facebook*, and blogs) in framing the Egyptian uprising of 2011. The authors found that the semi-official newspapers tended to highlight the conflict and chaotic side of the revolution, painting the protests as a devastating catastrophe and warning of economic crisis.

2.2. *Corpus linguistics and CDA*

Corpus linguistics is a branch of linguistics that consists of the study of language based on corpora as primary data, that is, machine-readable samples representative of authentic language use. It employs a quantitative and statistical methods of analysis for the empirical analysis of languages. It aims to find “probabilities, trends, patterns, co-occurrences of elements, features or groupings of features” in the corpora (Krishnamurthy and Teubert 2007: 6).

CDA is a qualitative approach based on the relationship between language and ideology (van Dijk 1997; Fairclough 2001; Wodak 2001). It describes, interprets, and explains the social inequalities in a discourse (Mullet 2018). Baker *et al.* (2008: 280) claimed that, in CDA, “language is not powerful on its own – it gains power by the use people make of it and by the people who have access to language means and public fora.” Hence, we used interdisciplinary work to analyse the language and understand how it transmits knowledge (Wodak 2004).

Therefore, the combination of these two approaches, CL and CDA, proves to be an effective method, as has been demonstrated by a plethora of language research. For example, it has been used to study the representation of Islam and Muslims in the British press (Baker 2012); the examination of the discourses of refugees and asylum seekers in the UK press (Baker *et al.* 2008); the representation of Gaddafi in Gulf countries’ media (Haider 2016); the analysis of discursive constructions of Scottish and British English national identities and their ideologies of independence in the press (Romero-Trillo and Cheshire 2014); and the examination of the Tunisian case after the Arab Spring (Romero-Trillo and Attia 2016).

This synergy provided us with CADS approach that we adopted to examine the language of the media in relation to the outcomes of the Libyan revolution. This approach aims to uncover meaning that might not be observed by the naked eye. As argued by Stubbs (1996: 92), “you cannot understand the world just by looking at it.” In this respect, we seek to expose the implicit bias in the media discourse by revealing the way in which war and terrorism issues were described during the post-revolution period. In this approach, we applied Bednarek and Caple’s (2017) DNVA framework.

2.3 The discursive news values analysis (DNVA)

Caple and Bednarek (2017) broadly define DNVA as a new approach to the analysis of news values that uses discourse analysis to examine how such values are constructed through semiotic resources (language, image, etc.).

We see news values as those values that have been considered in the literature as defining the newsworthiness of reported events and actors (where newsworthy means ‘worthy of being news’). This includes news values such as TIMELINESS, NEGATIVITY, IMPACT, SUPERLATIVENESS, ELITENESS, CONSONANCE and others listed in Table 1, which includes nine types of news values and how they are constructed as newsworthy, according to Caple and Bednarek (2017: 79).

News value	The event, including people, organisations, locations, etc. is (constructed as)
CONSONANCE	(Stereo)typical (for example, stereotypes regarding news actors, social groups, organisations or countries/nations)
ELITENESS	Of high status or fame
IMPACT	Having significant effects or consequences
NEGATIVITY	Negative (for example, a disaster, conflict, controversy, criminal act)
PERSONALISATION	Having a personal or ‘human’ face (involving ‘ordinary people’)
PROXIMITY	Geographically or culturally near the target audience
SUPERLATIVENESS	Of high intensity or large scope
TIMELINESS	Timely in relation to the publication date (new, recent, ongoing, about to happen, current, seasonal)
UNEXPECTEDNESS	Unexpected (for example, unusual, strange, rare)

Table 1: News values and their definitions (adapted from Bednarek and Caple 2017: 79)

Many linguists define news values from a linguistic perspective. Bell states that “these are *values*. They are not neutral, but reflect ideologies and priorities held in society” (Bell 1991: 156 [italics in original]). Fowler (1991: 13, 15) also claims that news values are not ‘natural’ but rather culturally and socially constructed. Richardson (2006: 93) mentions how news coverage focused on negative events such as war, terrorism, disaster and conflict in the developing countries, in a study by the Glasgow Media Group.

Despite the different analyses of the ideological aspects of news values on a wide range of topics (Bednarek 2016; Dahl and Fløttum 2017; Kitano 2019; Makki 2019, 2020), news values have not yet been the focus of critical linguistic analyses of news

discourse. Indeed, only a few studies have integrated DNVA with corpus techniques (Potts *et al.* 2015; Maklad 2019).

To bridge this gap, in this article, we aim to integrate DNVA in applying corpus techniques to reveal how the Libyan war events are linguistically constructed as newsworthy. Thus, we focus on the verbal semiotic resource. We use the term ‘event’ as a cover term for semiotic events, that is, issues and happenings. With an event’s value, we refer to the sociocultural side. In other words, we are interested in analysing how news values are constructed through language to establish the newsworthiness of the reported events (see further Bednarek and Caple 2012a, 2012b, 2014 and Caple *et al.* 2020).

This linguistic approach is followed by Potts *et al.* (2015) in a first case study on a large newspaper corpus on one culturally important event: the Hurricane Katrina. Their aim was to apply and test the integration of corpus techniques, in particular tagged lemma frequencies, collocation, key part-of-speech tags (POS-tags) and key semantic tags in DNVA. Similarly, Maklad (2019) adopted DNVA and a corpus linguistics approaches to explore how news values related to hate crime offenders and victims have been linguistically constructed in the American news media.

However, this approach of linguistic analysis applied only to English-language news, and it has just recently started to be applied to other languages such as Chinese (Caple *et al.* 2020), Spanish (Fuster-Márquez and Gregori-Signes 2019) and Persian/Farsi (Makki 2019). Yet, it has not been developed for the Arabic language. In this study, we will follow the strategy of Caple *et al.* (2020), who took ten relevant news values to the Chinese context (rather than the linguistic resources) as a departure point and inductively established whether they were constructed in the Chinese data. In our case, we selected six news values that are relevant to the Arabic news context and we later applied them to the English data: ELITENESS, IMPACT, NEGATIVITY, POSITIVITY, SUPERLATIVENESS and PERSONALISATION.

Thus, this study is the first to apply a DNVA framework to Arabic. For this reason, it is essential to take into account the target audience of each news category and to carefully analyse the construction of news values as they are context sensitive. However, it is also important to mention that this is not an exact science. It must be borne in mind that language is multifunctional, and thus, one text can have different interpretations. For example, one event can construct either ELITENESS or PERSONALISATION while another event can construct both NEGATIVITY and IMPACT at the same time.

3. METHOD

3.1. Corpus and procedure

The official online websites of each media outlet (*Al Jazeera* Arabic and English, *Al Arabiya* Arabic and English and CNN and BBC in English) were used to collect news reports that contained words related to Libya and the Libyan revolution and its outcomes. The search was filtered to December 2011–2015, as December 2011 marked the first sparks of the Arab Spring, and thus, the reported news from 2012 onwards during the last month of the year focused on covering and evaluating the news of the entire year.

The search code was the term *Libya*. This was selected to capture a large number of news reports that would be representative. The search was then filtered to show only news, and the focus was mainly on the local reporters of each media set rather than on the copies of the news agencies (like AP and Reuters). But, although some of the news could be taken from these news organisations, each news website has its own reporters to write or express the news to the reader. It can be the same piece of news, but it will be written in different styles. We made sure not to have any similarity in the copies of the different analysed media sets. Finally, we focused on the analysis of words rather than analysing the text as a whole.

After that, the relevant texts of the articles were manually extracted and converted to .txt. They were all manually checked to exclude irrelevant articles, like sport and art news. In all, our corpus stands at 413 news reports, collectively containing 174,804 words. Table 2 describes the origin and amount of the collected data.

Compared corpora	Number of articles	Number of words
<i>Al Jazeera</i> Arabic and <i>Al Arabiya</i> Arabic	174	55,205
<i>Al Jazeera</i> English and <i>Al Arabiya</i> English	165	67,912
BBC and CNN	74	51,687
Total	413	174,804

Table 2: Number of news reports and number of words in the corpus

The data were stored in three separate sub-corpora, as Table 2 shows, and inside each sub-corpus, five sub-corpora for the five different years (2011–2015) were included. The articles were then fed into *AntConc* 3.3.4 (Anthony 2018) for the analysis.

Our overall methodology is based on a corpus-assisted discourse analysis. Bednarek and Caple (2014) suggest that various corpus linguistic techniques can be used to study

newsworthiness. Thus, our focus was on the following corpus techniques: keywords, frequency lists, collocations and concordances. We adopted DNVA for the discourse analysis in order to analyse the topics of war and terrorism in reports of Libya after the revolution.

3.2. *Frequency analysis and keywords*

In this paper, the frequency analysis was generated for five years (2011–2015) in the three media categories named above. Partington (2006: 260) states that:

Corpus-assisted studies of register, genre or discourse type are of course by definition comparative: it is only possible to both uncover and evaluate the particular features of a discourse type by comparing it with others.

Therefore, we compared each year for each media category against the remaining years and media sets. The frequency lists were retrieved by means of the *WordList* tool in *AntConc* with the statistical measure (log-likelihood) and the keywords lists were generated through this comparison. This method was also followed by Garzone and Santulli (2004) when they compared the early responses of four British daily newspapers to the events of September 11. These lists were then compared through a manual analysis reading and based on the significance of the items that were attested in both the *Wordlist* and the keywords list. We selected items that suggest an interesting analysis and can uncover a number of ideological motifs. As a result, we found different topics that ranged from politics to economy, international relations, war and terrorism. However, we decided to focus on the topics of war and terrorism, due to the significant results obtained. Figure 1 shows an example of the wordlist and the keyword results extracted from *AntConc* for the English language Arab media in the year 2012. We highlighted some of the significant words attested in both lists that related to the topic of war; those are examples of words that are analysed in the present research.

antconc_results-wordlist - Bloc-notes					antconc_results-Keyword - Bloc-notes				
Fichier	Edition	Format	Affichage	Aide	Fichier	Edition	Format	Affichage	Aide
18	55		security		#Types Before Cut: 2311				
19	54		as		#Types After Cut: 99				
20	51		we		#Search Hits: 0				
21	48		has		1	38	60.310	clinton	
22	47		by		2	40	55.537	she	
23	46		an		3	32	44.614	department	
24	45		have		4	22	24.789	her	
25	45		state		5	44	23.480	benghazi	
26	44		benghazi		6	34	22.421	us	
27	44		from		7	35	21.985	attack	
28	44		who		8	45	21.859	state	
29	43		gaddafi		9	22	20.822	secretary	
30	42		had		10	15	17.783	diplomatic	
31	41		is		11	11	17.776	olive	
32	40		she		12	23	16.870	report	
33	39		were		13	10	16.160	assassination	
34	38		clinton		14	10	16.160	inadequate	
35	38		he		15	9	14.544	yunes	
36	37		after		16	14	14.158	th	
37	37		its		17	29	13.597	i	
38	37		which		18	11	13.143	my	
39	36		al		19	55	13.083	security	
40	36		but		20	14	12.926	place	
41	35		attack		21	9	12.519	issa	
42	35		be		22	9	12.519	obama	
43	34		been		23	8	11.874	levels	
44	34		his		24	7	11.312	failures	
45	34		us		25	7	11.312	grossly	
46	32		department		26	100	11.277	was	
47	31		libyan		27	15	11.126	ambassador	
48	29		i		28	109	10.696	xa	
49	28		new		29	7	10.266	gadaffi	
50	27		oil		30	8	10.065	posts	
51	25		it		31	51	9.765	we	
52	24		killed		32	6	9.696	accepted	

Figure 1: Wordlist and keyword list results of the English language (EL) Arab media in 2012

3.3. Collocation analysis

The term ‘collocation’ was coined by Firth (1957) and was defined by Baker (2006: 96) as “the phenomenon that certain words often co-occur with each other.” In other words, collocations are (groups of) words that are frequently attached to a target term, provide it with valuable information, and can also “convey messages implicitly” (Hunston 2002: 109). Stubbs (2001: 29) states that “software can calculate collocations by observing how many times the word x occurs near the word y.” This technique was frequently adopted in the analysis of our data, using the log-likelihood ratio to identify higher-frequency words, and it is preferred here to identify statistically significant collocates. This step was supported by manual analysis to exclude some irrelevant items. Thus, the most statistically significant collocates of our selected keyword lists across each sub-corpus are singled out for further analysis.

3.4. Concordance analysis

According to Freaake *et al.* (2010: 28):

A concordance search shows all the occurrences of the search term (or phrase) and its immediate co-text; concordance lines [can be] expanded to the whole text when needed.

Thus, in this paper, we examined some concordance lines with the help of the CDA approach to scrutinise with care the language of the studied media outlets and reveal their biased and violent language. Baker *et al.* (2008: 290) state that “[a] corpus-assisted approach, which looks for specific linguistic patterns and carries out tests of statistical significance is, therefore, able to quantify notions like ‘bias’.”

However, there were some problems when creating concordance lines for Arabic, since the software does not fully support right-to-left languages, which led to manual support for putting the words in the right order.

4. RESULTS

The categories war and terrorism are the area of scope in this article. They were examined year by year. The frequency distribution is expressed with raw frequencies (RF) and percentages (%) to yield valid results.

4.1. The year 2011

The frequency distribution analysis in the year 2011 highlights the similar percentage of reporting on the topic of war within the English language media sets, both Arab (1.28%) and Western (1.08%), but only a lower percentage (0.58%) within the Arabic Language (AL) editions of the Arab media, as shown in Table 3.

Categories	AL: Arab media		EL: Arab media		EL: Western media	
	RF	%	RF	%	RF	%
War	61	0.58%	262	1.28%	117	1.08%

Table 3: Frequency distribution of the category war in 2011

We started to analyse the keywords that refer to the Libyan national security, that is, *security, army, military, police* and *brigades*, as they appeared across our selected media categories. The elements represented by these keywords are considered high-status professionals, which generally construct ELITENESS, as illustrated in Table 4.

Media	Keyword	Frequency	Collocations
AL: Arab media	الأمن <i>Ālḥmn</i> ‘security’	21	واستعادة, والأمان, والاستقرار: Arabic terms; <i>wāst ‘ādā</i> ‘restore’, <i>wāl ḥmān</i> ‘security’, <i>wālāstqrār</i> ‘stability’: transliteration and translation
	<i>Army</i>	55	<i>uniting, structure, strong, strive,</i> <i>reinstated, reforming, oppressor,</i> <i>marginalised, inject, forming</i>
EL: Arab media	<i>Military</i>	34	<i>gain, effective, disputing, displayed</i>
	<i>Brigades</i>	16	<i>uniting, pledge</i>
EL: Western media	<i>Army</i>	15	<i>stockpiles, securing, professionalising</i>
	<i>Police</i>	14	<i>stockpiles, revived, professionalising</i>

Table 4: Frequency and collocations of ‘National security’ keywords in 2011

The collocations in the three media categories refer to the desire of the Libyan transitional government and national security forces to restore peace and regain stability and security. We can notice that the AL editions of the Arab media suggest that the transitional government worked hard to restore and strengthen stability and security in the country. The same results were attested in the EL editions of the Arab media; the elements represented by the keywords that construct the news value ELITENESS, *army*, *military* and *brigades*, formed a unity by making strong structures and radical reforms together. These high-status government positions were striving to reinstate security to the country and also pledged to restore peace. Similarly, the EL editions of the Western media suggested that national security, such as the police and the army, started to work together after the revolution to protect the country. The collocation *professionalising* highlights their professionalism to secure weapon stockpiles. Examples (1) and (2) below highlight the efforts of the transitional government to strengthen these forces.

- (1) [...] be handed over to the recently revived military **police**. But on the streets, young men in mismatched [...] (BBC 15/12/2011)
- (2) [...] until the government’s revived national army and **police** force are seen as strong enough to take [...] (BBC 15/12/2011)

The words *revived* and *professionalising* suggest that the transitional government in Libya was working hard to restore security and stability by strengthening the national forces. These results suggest that the collocations and the concordance findings are potential pointers to a specific news value, which is POSITIVITY, due to their constant encouragement to reunite the country. However, it cannot be denied that this event can construct a NEGATIVITY news value to a certain type of audience that may encourage the instability in Libya for a specific purpose such as, for example, to flee easily from the

country, to support the old regime of Gaddafi, etc. Moreover, it is important to mention that the keyword *police* in examples (1) and (2) above is generally associated with negativity; in Bednarek and Caple's words (2014: 8), "police deals with crime." Therefore, NEGATIVITY is more commonly considered as a news value than POSITIVITY (though see Schulz 1982: 152 and Harcup and O'Neill 2001: 279 on positive news); in fact, NEGATIVITY has been called "the basic news value" (Bell 1991: 156).

Indeed, the keyword results for 2011 suggest that many names of weapons were depicted in the analysis of the three media categories, together with other violent words. For example, the words *rebels*, *weapons*, *guns*, *missiles*, *war*, *crimes*, *fighters* and *killed* were frequently attested. These negative keywords describe negative events that were happening. The list of collocations in Table 5 emphasises this NEGATIVITY news value, which can be understood by all types of audiences.

Media	Keyword	Frequency	Collocations
AL: Arab media	السلاح <i>Ālslāḥ</i> 'weapon'	16	يريدون, تتدد, مناشدا, لافتات, الفوضى, انتشار Arabic terms; <i>yrydwn</i> 'want', <i>mdd</i> 'denounce', <i>mnāšdā</i> 'plead', <i>lāftāt</i> 'banners', <i>ālfwḍ</i> 'chaos', <i>āntšār</i> 'spread': transliteration and translation
	<i>Rebels</i>	39	<i>clashed, seats, killings, battled, autocratic</i>
	<i>Weapons</i>	41	<i>stripping, sweep, stores, experts,</i> <i>specialists, removed, rigged, pledge</i>
EL: Arab media	<i>Guns</i>	10	<i>propelled, mortars, machine, aircraft, haul</i>
	<i>Missiles</i>	7	<i>tank, stockpile, helicopter, assessing,</i> <i>armour, produce, mines, kilograms</i>
	<i>War</i>	39	<i>genocide, remnants, scapegoat, seize,</i> <i>recovering</i>
	<i>Crimes</i>	21	<i>humanity, unlawful, genocide, accuse,</i> <i>suspicious, committed, tribunal</i>
EL: Western media	<i>Rebel</i>	10	<i>dying, commanders, battle</i>
	<i>Killed</i>	22	<i>warehouse, trial, shots, missing, crossfire,</i> <i>allegations, battle, airstrike</i>
	<i>War</i>	20	<i>torn, terrible, sanctuary, overseas, civil,</i> <i>heed</i>
	<i>Guns</i>	10	<i>truck, startling, scaring, rifles, heavy,</i> <i>carrying, hands, reluctant</i>
	<i>Crimes</i>	7	<i>investigate, investigations, investigating,</i> <i>humanity, tyranny</i>

Table 5: Frequency and collocations of keywords for weapons names in 2011

Most of the collocations in the three media categories suggest a chaotic situation in Libya in 2011. The collocations indicate that fighters and rebels, who battled the autocratic

system that resulted in many killings, started some clashes with the national security forces in their struggle to gain seats in the National Transitional Council (NTC). These clashes made it difficult for the NTC and its forces to restore peace, especially with the spread of weapons and the wellspring recruitments by the rebels and their mobilisation. As a result, most verb and noun collocates clearly establish NEGATIVITY, through reference to the negative effects of the Libyan Revolution.

Although the specialists and experts pledged to sweep up all the weapons from the stores, crimes and genocides were still taking place. This created a mass militarisation that tipped the balance from a peaceful protest to an armed civil war (Bhardwaj 2012), as is highlighted by the collocations *shots*, *crossfire*, *airstrikes*, *dying*, *battle* and *insurgency*, which also construct NEGATIVITY.

This alarming situation can be corroborated by the African Union's Peace and Security Council meeting that took place in 2011, where the situation in Libya was defined as "a serious threat to peace and security in that country and the region as a whole" (Peace and Security Council 2011: 1). They further added

strong and unequivocal condemnation of the indiscriminate use of force and lethal weapons, whomever it comes from, resulting in the loss of life, both civilian and military, and the transformation of pacific demonstrations into an armed rebellion.¹ These quotes and collocations highlight Libya's ugliest underside in 2011 (2011: 1).

In this sense, we can conclude that 2011 was a massively hard year for the Libyans. From DNVA and corpus techniques perspectives, we can illustrate that the different media discourses share not only the same sense of negativity by showing a precarious and insecure situation, but also the same stance of positivity by beautifying and strengthening the image of Libya's national security forces in their attempt to restore peace amid unrest, which can also be negative to a specific type of audience.

4.2. *The year 2012*

The distribution in the year 2012 is almost the same as in the previous year. The topic of war is most frequent in the English language media sets (both Arab and Western), unlike the AL media, as can be seen in Table 6.

¹ AU document PSC/PR/COMM.2(CCLXV), 10 March 2011.

Categories	AL: Arab media		EL: Arab media		EL: Western media	
	RF	%	RF	%	RF	%
War	31	0.35%	109	1.09%	163	1.73%

Table 6: Frequency distribution of category war in 2012

In 2012 media results, the keyword *security* is attested in the three media categories in both languages, with different collocations and news values, as illustrated in Table 7.

Media	Keyword	Frequency	Collocations
AL: Arab media	الأمنية <i>Ālḥmnya</i> 'security'	16	Arabic: السلاح, المخاوف, قدرات, ضعف, وفوضى terms; <i>wfwḍa</i> 'chaos', <i>d'f</i> 'weakness', <i>qdrāt</i> 'abilities', <i>ālmḥāwf</i> 'fears', <i>āslāḥ</i> 'weapons': transliteration and translation
EL: Arab media	<i>Security</i>	55	<i>strengthen, risks, restored, progress</i>
EL: Western media	<i>Security</i>	80	<i>strengthening, solving, resolution</i>

Table 7: Frequency and collocations of the keyword *security* in 2012

Although the term *security* should construct POSITIVITY as it refers to terms such as *safety*, *protection*, *certainty*, etc., it constructs NEGATIVITY in the AL editions of the Arab media. It is related to negative terms like *weapons*, *chaos*, *weakness* and *fears*, which highlight the fragility of that period in Libya. Moreover, although *security* points to POSITIVITY in the EL editions of the Arab media, such as *strengthen*, *restored* and *progress*, this hypothesis was not supported by an analysis of the concordances, as shown in examples (3) and (4) below. These examples include pointers to NEGATIVITY, such as *inadequate*, *failures* and *failings*, which suggest the failure of the country to provide citizens with security and safety.

- (3) [...] State Department resulted in a Special Mission **security** posture that was inadequate for Benghazi [...] (*Al Jazeera* 19/12/2012)
- (4) with Republicans skewering the administration for **security** failings as well as a possible cover... (*Al Jazeera* 19/12/2012)

Similarly, in the EL articles of the Western media, the high frequency of the term *security* (80 tokens) denotes the persistent attempt of the government to restore peace, stability and security, as the collocates lend credence, namely *strengthening*, *solving* and *resolution*. However, the concordances also show similar results to those attested in the EL editions of the Arab media, as illustrated in (5) to (7) below.

- (5) [...] independent report that found “grossly inadequate **security**” and “systemic failures and [...] (CNN 28/12/2012)

- (6) [...] military intervention is essential to solving the **security** crisis. When soldiers seized the capital [...] (CNN 12/12/2012)
- (7) [...] Security and Government Affairs, cites “extremely poor **security** in a threat environment that was ‘flashing red.’ (CNN 31/12/2012)

Examples (5) to (7) pinpoint NEGATIVITY, IMPACT and SUPERLATIVENESS as news values, which are emphasised through several expressions from the concordance analysis: *grossly inadequate security*, *security crisis* and *extremely poor security*. These results indicate not only the fragility of the security situation in Libya, but also the ideological bias of the Western media that is trying to represent a failed country with an ineffective security system. Indeed, the negative ideology was shared across the three media categories.

Moreover, the collocations in Table 8 construct different news values of Libya’s security situation.

Media	Keyword	Frequency	Collocations
AL: Arab media	الشرطة <i>Ālṣrṭ</i> ‘police’	15	القنبلة، بتفجيرها، وهدد، استسلامه، تفاوضت Arabic terms; <i>tfāwḍt</i> ‘negotiated’, <i>āstslāmh</i> ‘surrendered’, <i>whdd</i> ‘threatened’, <i>bṭḡyrhā</i> ‘detonated’, <i>ālqnbḷ</i> ‘bomb’: transliteration and translation
	<i>Attack</i>	35	<i>kill, lapse, fallout, deadly, Christopher</i>
EL: Arab media	<i>Assassination</i>	10	<i>torture, betrayal, Yunes</i>
	<i>Corruption</i>	9	<i>struggling, growing, anger, complained</i>
EL: Western media	<i>Attack</i>	54	<i>unconscious, terrible, surprising, sudden, opportunistic, enemy</i>
	<i>Militants</i>	12	<i>strength, growing, attacked, warning, strict, Islamist, armed, trick</i>
	<i>Rebels</i>	9	<i>Tuareg, ethnic, deployment, rebellions, staged, fighting, desert, battle</i>
	<i>Deadly</i>	8	<i>attacks, attack, Benghazi</i>

Table 8: Frequency and collocations of keywords related to *attacks* in 2012

The keywords *attack*, *assassination*, *deadly* and *corruption* attested in the three media categories construct NEGATIVITY. Moreover, the collocation *Christopher* constructs IMPACT, as it refers to the US ambassador in Benghazi consulate Christopher Stevens, who was attacked and killed. The Western media described his assassination as *sudden*, *surprising* and *terrible*, which are pointers to SUPERLATIVENESS and NEGATIVITY. These news values make the lapse of security in Libya explicit. This event can, however, construct POSITIVITY to a specific audience, who can be either encouraging violence and

chaos in Libya for their own benefits or hating the intervention of the US in the local issues of Libya. In the same vein, the collocations, *struggling*, *anger* and *complained* in the EL editions of the Arab media construct IMPACT, as they denote the spread and growth of corruption in Libya, plus the popular outrage that evolved into a civil war.

Moreover, the keyword *militants* in the EL coverage of the Western media constructs ELITENESS, as it refers to a high-weight group in Libya, that is, the Islamist militants. It also constructs NEGATIVITY, which is proved by the collocations *strong*, *strict*, *tricky* and *armed*. These collocations suggest the strength and danger of this Islamist group. The same results are found for the keyword *rebels*, which is assigned to *Touareg*, a specific ethnic category in Libya. The Touareg were rebellious for decades, and they took advantage of the power vacuum in Libya to fight and seize some areas. However, this event can also construct POSITIVITY to the audience that supports Touareg and the Islamist groups, as chaos may give them more power and strength to set their camps.

To summarise, the news values analysis with corpus techniques shows that the three media categories shared the same story content and negative ideology, exemplified by the NEGATIVITY news value towards Libyan security and the country's woeful situation in 2012.

4.3. The year 2013

A new topic burst in relation to Libya in the findings for the year 2013, that is, terrorism; however, it is not attested in the EL editions of the Arab media. Moreover, the Western media has the highest percentage of references to the topics of war and terrorism in Libya, as Table 9 shows.

Categories	AL: Arab media		EL: Arab media		EL: Western media	
	RF	%	RF	%	RF	%
War	33	0.29%	61	0.69%	100	1.84%
Terrorism	31	0.27%	0	0.00%	60	1.10%

Table 9: Frequency distribution of the categories of war and terrorism in 2013

4.3.1. Topic of war

As was the case in 2011 and 2012, the findings for the year 2013 also reveal the spread

of militias and tribal rebellions all over Libya that caused a state of terror. The collocations in Table 10 clearly construct NEGATIVITY, IMPACT and PERSONALISATION in the Libyan situation.

Media	Keyword	Frequency	Collocations
AL: Arab Media	الاغتيالات <i>Alāḡtyālāt</i> 'assassinations'	14	مصادمات, انفلات: Arabic terms; <i>ānflāt</i> 'breakdown', <i>mṣādmāt</i> 'clashes': transliteration and translation
	قتل <i>Qtl</i> 'kill'	10	اندلاع, هجوم, أصيب, ضابطان, ضحيته, اغتيالات: Arabic terms; <i>āḡtyālāt</i> 'assassinations', <i>ḡhythā</i> 'victim', <i>dābṭān</i> 'two officers', <i>ḡyb</i> 'wounded', <i>hḡwm</i> 'attack', <i>āndlā</i> 'breakout': transliteration and translation
	مجهولين <i>Mḡhwlyn</i> 'unknown people'	9	هجومين, برصاص, خطيرة, شرطيا, مسلحين: Arabic terms; <i>mshlyn</i> 'gunmen', <i>ṣrṭyā</i> 'policeman', <i>hṭyrā</i> 'dangerous', <i>brṣāṣ</i> 'shot', <i>hḡwmyn</i> 'two attacks', <i>āḡtālāwā</i> 'assassinated': transliteration and translation
EL: Arab Media	<i>Militias</i>	28	<i>shots, risk, rein, resisting, killings, tribesmen, servants</i>
	<i>Bombing</i>	12	<i>shooting, rattle, explosive, airliner, suicide, mourning, convicted</i>
	<i>Tribesmen</i>	11	<i>servants, minorities, civil, workers, militias, seized, oilfields, rein, problem, political</i>
	<i>Shot</i>	10	<i>dead, Mosbah, teacher, colonel, gunmen</i>
EL: Western Media	<i>Smith/teacher/Ronnie/American</i>	28/21/ 10/14	<i>threatening, slain, murdered, shooting, gunmen, attackers, blood, shot, death</i>
	<i>Violence</i>	9	<i>kills, gripped, chaos, bursts, militia, assassinations, bombings, political</i>
	<i>Killing</i>	8	<i>investigate, diplomatic, ambassador, Americans, fire</i>
	<i>Assassinations</i>	5	<i>chaos, political, violence, ambassador, spiked</i>
	<i>Kidnapping</i>	5	<i>shooting, capture, revenge, Americans</i>

Table 10: Frequency and collocations of violence-related keywords in 2013

The three media categories reported on the events that caused an armed civil war, as shown in Table 10. Some examples were selected to highlight these events and their news values. We selected first some collocations from the three media categories that construct NEGATIVITY and IMPACT, and then we chose some collocations that construct PERSONALISATION:

- AL: Arab media: *assassinations, clashes, breakdowns, attacks, shots*, etc. (NEGATIVITY and IMPACT).
- EL: Arab media: *shooting, explosive, suicide, mourning, rattle, convicted*, etc. (NEGATIVITY and IMPACT).
- EL: Western media: *bursts, bombings, violence, chaos, shooting, killing, assassinations, kidnapping*, etc. (NEGATIVITY and IMPACT), but also, many *assassinations*.
- AL: Arab media: *two officers, police, victim* (PERSONALISATION).
- EL: Arab media: *the army colonel* (Fethallah al-Gaziri), *the American teacher: Ronnie Smith* and *Mosbah el-Kabaeli* (PERSONALISATION).
- EL: Western media: *the American teacher: Ronnie Smith* (PERSONALISATION)

Thus, we observe that the three media categories portrayed Libya in an alarming situation, rife with assassinations and kidnappings, that were exemplified by NEGATIVITY, IMPACT and PERSONALISATION of the victims. The results elucidate a state of horror. The story content and the negative ideology are the same throughout these media.

However, as a general fact, any fight or war has supporters and adversaries. That is why the results that suggest Libya to be a state of horror can construct POSITIVITY to those supporters of the Libya's chaos in order to promote their own interests.

4.3.2. Topic of terrorism

The term *Shariaa* in the AL editions of the Arab media and the EL reporting of the Western media refers to *Ansar al-Shariaa* ('supporters of Islamic Law'). This term is a pointer to ELITENESS, as it is considered a Salafist Islamist militia group that advocated the implementation of an extreme Islamic law across Libya (Irshaid 2014). This organisation is considered a terrorist movement by many powerful countries, such as the United Kingdom, the United States, the United Arab Emirates and Turkey (Home Office 2016).

The collocations that are associated with this movement in the AL editions of the Arab media (*debate, acknowledge, notes* and *confirms*) strongly elucidate the dominant position that Ansar al-Shariaa has in making decisions in Libya (see Table 11).

Media	Keyword	Frequency	Collocations
AL: Arab Media	الشريعة <i>Ālšry</i> 'ة' 'Shariaa'	31	يؤكد, يشر, يقر, والنقاش: Arab terms; <i>wālnqāš</i> 'debates', <i>yqr</i> 'acknowledges', <i>yšr</i> 'notes', <i>ykd</i> 'confirms': transliteration and translation
	<i>Qaeda</i>	38	<i>sympathisers, supporters, sympathetic, planner, scary</i>
EL: Western Media	<i>Sharia/Ansar</i>	8/5	<i>Islamic, Islamist, militant, militia, deadly, political, assassinations</i>
	<i>Islamist</i>	9	<i>militiamen, hardliners, militant, tortured, Islamists, militia, arrested, anger</i>

Table 11: Frequency and collocations of keywords related to *terrorist groups* in 2013

Therefore, the results from the AL editions construct UNEXPECTEDNESS, which highlights the powerful and controlling position of Ansar al-Shariaa. However, the collocations in the EL articles of the Western media construct NEGATIVITY, since they highlight its association with terrorist words, such as *assassinations, deadly, militant* and *militia*. These violent words are also related to *Islamic* and *Islamist*, which seems a suggestion from the Western media that Islam is connected to terrorism. Similarly, the salient term, *Islamist* is associated with violent words like *tortured, arrested* and *anger*, as well as with *militiamen, militia, militant* and *hardliners*, which are pointers to NEGATIVITY.

Moreover, the EL coverage of the Western media reported on another group called Al Qaeda, which is a militant Sunni Islamist organisation (Bergen 2006). It has been listed as a terrorist movement by the United Nations Security Council, the North Atlantic Treaty Organization (NATO), the European Union and several other countries. It is also a multinational movement supported by Islamist extremists from all over the world. As shown in Table 11, the collocations *sympathisers* and *supporters* highlight the firm position that Al Qaeda had in Libya by gaining support and sympathy from the different militia groups that joined the movement. However, from the collocation *scary* it can be suggested that civilians fear this movement. Therefore, these collocations can construct NEGATIVITY to the audiences that do neither agree with nor support Al Qaeda and can construct POSITIVITY at the same time to the audiences that sympathise and encourage the principals of this organisation.

It can, therefore, be argued that the AL editions of the Arab media and the EL coverage of the Western media construct the same NEGATIVITY towards the militia groups that arose in Libya in 2013 and their dominant position in the country.

4.4. The year 2014

Both categories of Arab media covered the topic of war in 2014 with percentages similar to those attested in previous years. However, the topic of terrorism was covered only by the EL editions of the Arab media with a low percentage (see Table 12).

Categories	AL: Arab media		EL: Arab media		EL: Western media	
	RF	%	RF	%	RF	%
War	228	1.51%	266	1.77%	0	0.00%
Terrorism	0	0.00%	40	0.26%	0	0.00%

Table 12: Frequency distribution of the categories war and terrorism in 2014

4.4.1. Topic of war

The findings for the AL editions of the Arab media show the attestation of new war factions that are pointers to ELITENESS, since they have a high value in the country, *Fajr Libya* ('Libya Dawn') and *Haftar's army group*, and the emergence of two rival governments. The collocations and the salient words suggest (see Table 13) the development of a civil war between Haftar's group, Fajr Libya, Libya Shield Force and other militias seeking control of the territory and oil of Libya.

Keyword	Frequency	Collocations
فجر <i>Fġr</i> 'fajr'	92	مسلحي, مقاتلين, انصار, باشتباكات, وحرس, درع, كتائب, ميليشيا, ميليشيات Arabic terms; <i>mylyšyāt</i> 'militias', <i>mylyšyā</i> militia', <i>ktāšb</i> 'battalions', <i>dr</i> 'shield', <i>wḥrs</i> 'guards', <i>bāštbākāt</i> 'clashes', <i>ānšār</i> 'ansar', <i>mḡātlyn</i> 'fighters', <i>mslḥy</i> 'militants': transliteration and translation
الجيش <i>ālġyš</i> 'the army'	54	Arabic terms; <i>yḥwd</i> 'fight', <i>syqsf</i> 'will bomb', <i>slāḥ</i> 'weapon', <i>wdbābāt</i> 'tanks': transliteration and translation
حفتر <i>hftr</i> 'Haftar'	42	Arabic terms; <i>yqwd</i> 'leading', <i>yšnhā</i> 'launching', <i>ṭrd</i> expelling', <i>āštbākāt</i> 'clashes', <i>hḡwmh</i> 'attacking': transliteration and translation
الميليشيات <i>ālmlyšyāt</i> 'militias'	21	Arabic terms; <i>m'skrāt</i> 'camps', <i>qā'dš</i> 'stronghold', <i>nšbt</i> 'put', <i>m'rktḥ</i> 'battle': transliteration and translation
اشتباكات <i>āštbākāt</i> 'clashes'	19	Arabic terms; <i>llḥkwmtyn</i> 'two governments', <i>nyf š</i> 'violent', <i>sqṭwā</i> 'fell', <i>ḡrḥ</i> 'wounded', <i>bḡrwḥ</i> 'injured', <i>ālmtnāfstyn</i> 'rivals', <i>āndl</i> 't' 'broke out': transliteration and translation

Table 13: Frequency and collocations of the topic of war in the AL editions of the Arab media in 2014

Fajr Libya militia, which is a non-jihadist group, is collocated with war and conflict terms, such as *militia, clashes, fighters, battalions* and *militants*, that point to both NEGATIVITY and IMPACT due to their continuous association with attacks and clashes. Moreover, the salient words *the army* and *Haftar* refer to the General Khalifa Haftar group (ELITENESS), a pro-government movement. Haftar is also famous for its hatred of the jihadist groups. The collocations construct, again, NEGATIVITY and IMPACT as they describe the launch by Haftar's armed group and the Libyan National Army (LNA) of a series of clashes and attacks, using bombs, weapons and tanks against Fajr Libya militia and others. Similarly, the other militias were bombed, which made it a fierce and furious battle. The clashes and bombardments also included the two governments that are operating in Libya; the Tobruk government, known as the House of Representatives, led by the Prime Minister Abdullah al-Thinni and backed by Haftar, and its rival, the Tripoli government, known as the General National Congress and backed by Fajr Libya (Libya Dawn).

Similarly, the high frequency of the keywords in the EL editions of the Arab media can only reinforce the impression of the catastrophic situation of Libya. The pointers of NEGATIVITY and IMPACT (*fighting, clashes, fire, violence, strikes*, etc.) describe again the painful continuum of chaos and violence in Libya. Moreover, the pointers of ELITENESS (*Fajr, Haftar, soldiers, Touareg* and *tribes*) portray the different factions in the country that are seeking to gain land, oil and ports, due to the lack of governance, which turned the situation into a civil war.

As shown in Table 14, the collocations related to the words of brutality and destruction that also construct NEGATIVITY, (*tumultuous, raged, crush, fighter, deadlock, fended, damaged, destruction, grave*, etc.) highlight the fighting and clashes between these different groups, which led not only to the escaping of many civilians to Tunisia and Algeria through borders but also to the damaging of the oil ports by fire due to airstrikes. These references construct the Impact of the Libyan situation.

Keyword	Frequency	Collocations
<i>Fighting</i>	35	<i>tumultuous, struggling, raged, escaped, escalated, crush</i>
<i>Fajr</i>	32	<i>dismantled, fighter, militiamen, radicals</i>
<i>Haftar</i>	31	<i>crush, waging, supporters, confrontations</i>
<i>Clashes</i>	28	<i>persist, fighter, fended, deadlock, broke</i>
<i>Border</i>	24	<i>Algeria, Tunisia, Tunisian, patrolling, destroying, closes, ripe, smuggling</i>
<i>Soldiers</i>	20	<i>kill, fighter, damaged, battalion, broke, die</i>
<i>Fire</i>	20	<i>spreads, extinguish, blaze, firefighters, damaged, tank</i>
<i>Tuareg</i>	18	<i>refused, supported, stormed, shifting, rebellion, indigenous, forge, tebu</i>
<i>Violence</i>	18	<i>triggering, targeting, trajectories, infighting, entangled, embroiled, destruction, deadlock, ending, grave</i>
<i>Strikes</i>	18	<i>flew, footage, aviation, targets, air</i>
<i>Terrorism</i>	13	<i>Westerners, strengthening, Islamism, threats, jihadist</i>
<i>Tribes</i>	9	<i>violently, pitted, shifting, nomadic, broken, allegiances</i>

Table 14: Frequency and collocations of *war* in the EL editions of the Arab media in 2014

This situation was caused by the spread of different groups across the country: *Fajr Libya*, *Haftar group*, *Touareg*, *tribes*, *soldiers*, *terrorists*, and many more. However, the new groups that appeared in 2014 were the extremist militants or terrorists that led the situation to fester, affecting not only Libya but all the bordering countries, which represented a greater threat to the west.

To summarise, the Arab media in both languages highlight the violent and bloody civil war that circulated in Libya. They both reported on the fighting among the different militia groups (NEGATIVITY, IMPACT and ELITENESS).

Yet, it is important to mention that this chaos and violent situation in Libya led to the birth of terrorists inside the country, which could be positive to their supporters.

4.4.2. Topic of terrorism

Similar to the EL coverage of the Western media in 2013, the topic of terrorism is related to the religion of Islam, which can be seen from the salient word *Islamist* in Table 15.

Keyword	Frequency	Collocations
<i>Islamist</i>	40	<i>kill, clashed, broke, threatening, smuggling</i>

Table 15: Frequency and collocations of the category terrorism in the EL editions of the Arab media in 2014

The word refers to the new groups of Jihadists or Islamist militants in Libya, as can be seen in examples (8) to (11).

- (8) At least 19 Libyan soldiers were killed in clashes with **Islamist** militia on Thursday east of the country [...] (*Al Arabiya* 25/12/2014)
- (9) [...] of Tripoli, is in the hands of **Islamist** militias including Ansar al-Sharia, which [...] (*Al Arabiya* 26/12/2014)
- (10)[...] following clashes between pro-government forces and **Islamist** militias, officials said on Friday. (*Al Arabiya* 26/12/2014)
- (11)[...] and the eastern city of Benghazi from **Islamist** militants. The Islamic State of Iraq and Syria [...] (*Al Arabiya* 13/12/2014)

It cannot be denied that the term *Islamist* constructs NEGATIVITY and IMPACT, through its association with *killed, clashes, Ansar al-Sharia, the Islamic State of Iraq and Sham*. It clearly reveals that the media emphasised the extent of violence, making terrorism appear to be the norm. Therefore, this media category suggests that the new group, Islamic State of Iraq and Sham (ISIS), is related to Islam and Muslims, which serve its agenda.

4.5. The year 2015

The year 2015 was very exceptional in Libya, as we can see from Table 16. The topic of war was no longer attested, and the focus switched to the topic of terrorism, widely attested in all the media, thus illustrating its negative effect on the country.

Categories	AL: Arab media		EL: Arab media		EL: Western media	
	RF	%	RF	%	RF	%
Terrorism	151	1.54%	164	1.17%	265	1.24%

Table 16: Frequency distribution of the category terrorism in 2015

In 2015, a broad range of leaders worked feverishly to erode the civil war, strengthen unity and establish democracy. As daunting as this task may seem, further difficulties could be seen on the horizon, due to the so-called ISIS.

The appearance of Al-Qaeda as a salient word in the AL editions of the Arab media category can only indicate its relationship to the Islamic State (IS). Indeed, they are both internationally considered terrorist movements, and they both construct ELITENESS.

Keyword	Frequency	Collocations
القاعدة <i>Ālqā‘d</i> ‘Qaeda’	47	Arabic: القيادة, الكتيبة, تنظيمي, محتدم, الصراع, وداعش, يخشى terms: <i>yḥš</i> ‘fearful’, <i>yṯḥwf</i> ‘afraid’, <i>wdā‘š</i> ‘Daesh’, <i>ālšrā‘</i> ‘conflict’, <i>mḥtdm</i> ‘franging’, <i>tnzymy</i> ‘organisational’, <i>ālkyb</i> ‘battalion’ <i>ālqyād</i> ‘leadership’: transliteration and translation
تنظيم <i>Tnzym</i> ‘organisation’ داعش <i>dā‘š</i> ‘ISIS’	43/36	Arabic terms: سحق, يتخوف, يتمدد, يسيطر, يرسل, يضم, ينفذها <i>ynfdhā</i> ‘implements it’, <i>yḍm</i> ‘combines’, <i>yrsł</i> ‘sends’, <i>ysyṯr</i> ‘controls’, <i>ytmdd</i> ‘expands’, <i>yṯḥwf</i> ‘fears’, <i>šḥq</i> ‘crushed’: transliteration and translation
الإرهاب <i>Ālḥrāb</i> ‘terrorism’	19	Arabic terms: لتخليص, لتطهيرها, للتعاون, لمحاربة, مكافحة, المساعدة <i>almsā‘d</i> ‘help’, <i>mkāfḥ</i> ‘strive’, <i>lmḥārb</i> ‘fight’, <i>llt‘āwn</i> ‘cooperate’, <i>ltṯhyrhā</i> ‘to cleanse it’, <i>ltḥlyṯ</i> ‘to get rid of’: transliteration and translation
مقاتل <i>Mqātl</i> ‘fighter’	6	Arabic terms: عديد, قاتلوا, آلاف, ألفي, ثمانمئة, وثلاثة <i>wṯlāt</i> ‘three’, <i>tmānm</i> ‘eight hundred’, <i>ḥfyn</i> ‘two thousand’, <i>ḥfy</i> ‘two thousand’, <i>ḥāf</i> ‘thousands’, <i>qātlwā</i> ‘fought’, <i>dyd</i> ‘many’: transliteration and translation

Table 17: Frequency and collocations of the category terrorism in the Arabic language in 2015

The collocations *organisational*, *battalion*, *leadership* and *Daesh* prove that these two extremist groups share the same category, organising conflicts and attacks together, and implementing fear in Libya and the world in general, which classifies them in NEGATIVITY and IMPACT news values. Yet, these events can construct POSITIVITY to their supporters, who are encouraging them to gain more control and expand more in the territory. As ISIS expanded in Libya, it succeeded in gaining control over many territories due to its large number of fighters, which can be proven by the collocations of the term *fighter*, as shown in Table 17: *three*, *eight hundred*, *two thousand* and *many*. These pointers to SUPERLATIVENESS focus on the number of fighters, which can be interpreted as an attempt to panic a certain type of audience, while an encouragement and pride to the supporters of this movement. However, the collocations related to the salient word *terrorism* show that the government strives to crush them and cleanse the country, which constructs SUPERLATIVENESS as well.

Likewise, the EL editions of the Arab media covered the new-born terrorist group, ISIS. The security vacuum in Libya allowed this group to gain a foothold. As shown in Table 18, the collocations *stolen*, *slain*, *perpetrate*, *militancy* and *massacred*, which construct IMPACT, show their terror attacks and their horrible massacres in the country. The salient words, *chaos*, *threat* and *militant* manifest the chaotic situation of Libya and the violent role of ISIS, pointing to NEGATIVITY and IMPACT with collocations such as *thrown*, *killing*, *ouster*, *violence*, *threatened*, *outrages*, *lawless*, *criminals*, *terror* and

exploiting. Moreover, the collocations of the salient word, *rival* indicate the struggle between the ELITENESS pointers (*administrations, governments and parliaments*) and the terrorist group ISIS that is obstructing them in their efforts to move the country forward.

Keyword	Frequency	Collocations
<i>ISIS /ISIL</i>	72	<i>warns, tackle, stolen, slain, perpetrate, militancy, massacred</i>
<i>Rival</i>	48	<i>wracked, obstructing, impress, administrations, governments, parliaments</i>
<i>Chaos</i>	23	<i>thrown, sunk, preventing, killing, ouster, rebellion, violence</i>
<i>Threat</i>	14	<i>threatened, outrages, lawless, grapple</i>
<i>Militant</i>	7	<i>criminals, extremist, terror, struggling, exploiting</i>

Table 18: Frequency and collocations of *terrorism* in the EL Arab media in 2015

Similarly, the findings of the EL coverage of Libya in the Western media category indicated the spread and danger of ISIS in Libya and worldwide. As illustrated in Table 19, the collocations of the salient words *ISIS, attacks, terrorism* and *chaos* demonstrate the successful attempt of ISIS to take over some territories.

Keyword	Frequency	Collocations
<i>ISIS</i>	164	<i>shutting, punishments, warns, volunteers, takeover</i>
<i>Attacks</i>	35	<i>threaten, struggled, Milan</i>
<i>Rival</i>	28	<i>tribes, politicians, lawmakers, fuelled, governments, parliaments, militia</i>
<i>Terrorism</i>	27	<i>suffer, Pakistan, Sinai, protecting, hotbed, homegrown</i>
<i>Chaos</i>	11	<i>overthrow, Muammar, removal, revolution, instability</i>

Table 19: Frequency and collocations of *terrorism* in the Western media in 2015

Moreover, it cannot be denied that this group spread around the globe, as indicated by the frequent collocation of the word *attacks* with placenames, such as *Nigeria, Mali, Egypt, US* and *Europe*, which construct PROXIMITY (see (12) to (15)).

(12)[...] up key national security nominations.” Terrorist **attacks** in San Bernardino, California, and Paris [...] (CNN 16/12/2015)

(13)[...] members still to continue audacious and deadly **attacks**. Nigeria cannot afford to ignore the large [...] (BBC 21/12/2015)

(14)[...] blamed on Tuareg and Islamist groups 2015: Terror **attacks** in the capital, Bamako, and central Mali [...] (BBC 21/12/2015)

(15)[...] both the scale and frequency of its **attacks** on the Egyptian military have grown exponentially. (CNN 31/12/2015)

Despite this extremist group's terror attacks all over the world, all the Libyan parties (ELITENESS), that is, tribes, politicians, lawmakers, governments and parliaments, gathered and united to combat them and restore peace, stability and security in Libya.

The results of the analysis of the media reports in 2015 highlight the bias of the media when reporting on the outcomes of the Libyan revolution. Although the government was working hard to restore peace and security, the media focused mainly on the attacks of the new militia group, ISIS, as proved by ELITENESS, IMPACT and NEGATIVITY news values.

5. DISCUSSION

This cross-linguistic study explored, through DNVA and corpus techniques, a new dimension on how some news websites covered Libyan news events, specifically the topics of war and terrorism, during the post-revolutionary period.

To answer the first research question, namely, whether there is any significant difference in terms of distributions or lexical frequencies between the three categories of news articles, we found that the quantitative analysis of the frequency distribution highlights the similarities and differences between the three media categories under investigation. As we can see in Figures 2, 3 and 4 below, the category war had low percentages (AL 0.29% and EL 0.69%) in the two Arab media editions in 2013, while the Western media showed a higher frequency (1.84%). In contrast, in 2014, the Western media showed a remarkable drop of the frequency of the category war, while there was a high frequency of this category in the two Arab media editions. However, the topic of terrorism showed a sudden burst all in the media categories during 2015.

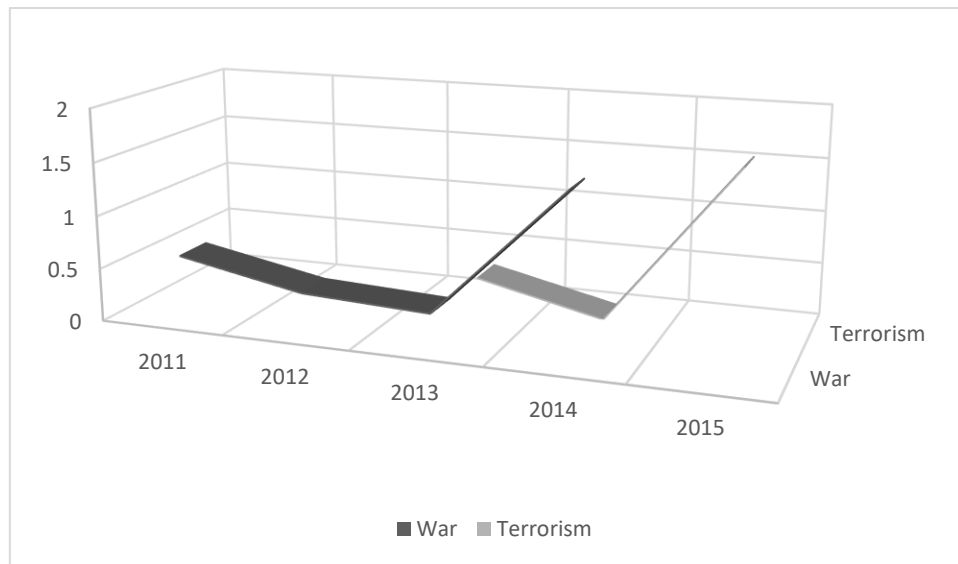


Figure 2: Frequency distribution of the categories war and terrorism in percentages in the AL editions of the Arab media

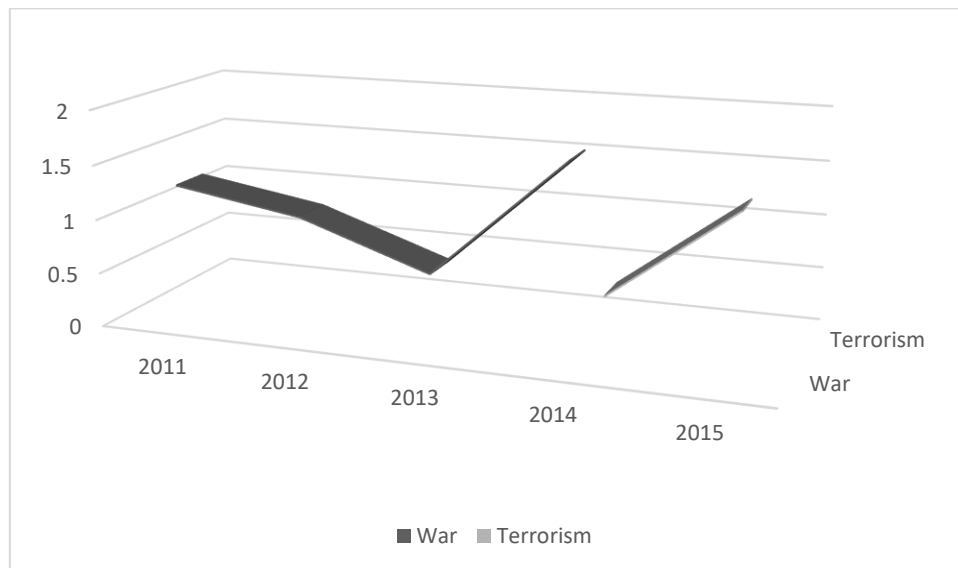


Figure 3: Frequency distribution of the categories war and terrorism in percentages in the EL editions of the Arab media

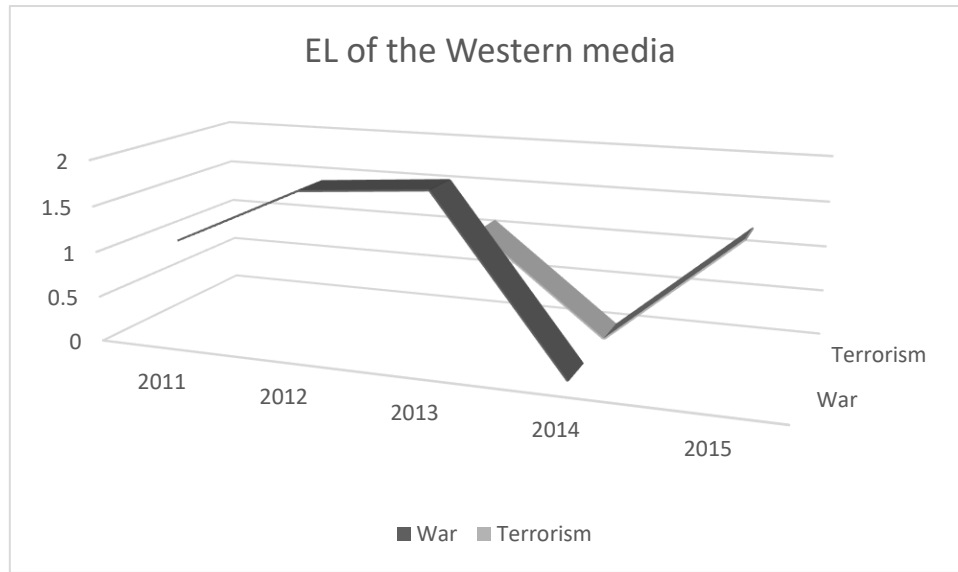


Figure 4: Frequency distribution of the categories war and terrorism in percentages in the EL editions of the Western media

As for the second research question, which enquires whether the salient words reveal the same story contents within the three analysed categories of media, the results show that the three media types have the same story contents in the categories of war and terrorism during the period under investigation. The quantitative and qualitative analyses confirm the use of the same conflict discourse in the AL and EL editions of the Arab media and the EL output of the Western media, despite their different ideological backgrounds, which, in turn, provides an answer to the third research question which deals with the role of collocations, concordances and DNVA in the agendas and ideological orientations of the media organisations. Therefore, when it comes to a situation such as that of Libya, which has the focus of the media, it becomes the wellspring of conflicts and battles and the arena of civil war and terrorism. In this context, the language of violence is assumed in the discourse of all the media dealt with in the present study. The high frequency of the news value NEGATIVITY proves that the war became the mainstay of the media, as it was persistent during the four years following the revolution (2011–2014) and this category reflects the terminology of battles, clashes, killings and weapons. Therefore, the DNVA approach proves that all the media employed violent discourse when describing the situation of Libya during these years.

Nevertheless, further analysis of the collocations and the concordances in all the years under analysis here proves that this positive perspective cannot cover up the violent orientation of the media. Indeed, violent terms, such as *shots*, *crossfire*, *dying*, *battle*, *assassinations* and *bombings*, are frequently used. As a result, these media share the same

negative and violent story contents and ideologies during the five years analysed (2011–2015), which aligns with the view of Marsden and Savigny (2010) on how the media can play a significant role in shaping political agendas and easily manipulate the audience.

As regards our fourth research question, which addresses what news values are discursively constructed in different languages, the Arab media repeatedly portrayed a perilous image of the terrorist organisations that constructed ELITENESS (Ansar Al-Shariah and Al Qaeda) and the new-born extremist movement: ISIS. They were collocated with NEGATIVITY news values terms, such as *killing, assassinations, violence, threatened, criminals* and *terror*. This persistent negative tone of coverage within the Arab media was also attested in Fahmy and Emad (2011), and in Romero-Trillo and Attia (2016) for the reporting of Tunisia in the Western media.

The findings indicate the danger of ISIS, which not only gained some territory in Libya but also spread worldwide, as shown by some concordances that constructed PROXIMITY, such as *Nigeria, Mali, Egypt, US, Europe* and *Milan*. Moreover, the results of the story content of the English language media (both Arab and Western) associate terrorism with Islam. For example, they include references to Islam in association with negative expressions of killings and threats that also construct NEGATIVITY, such as *kill, clashed, broke* and *threatening*. The attempt to relate terrorism to Islam in the media was confirmed by Törnberg and Törnberg (2016).

In sum, the frequency analysis shows remarkable similarities and differences in the coverage of war and terrorism in the three media categories. As regards the story content, the salient words show that the three media have the same story contents in the categories of war and terrorism during the period under study. This is shown in the analysis of the news values, which mainly construct NEGATIVITY, IMPACT and ELITENESS, highlighting the negative and violent discourse of the three media, and, in turn, shaping the readers' beliefs. However, PERSONALISATION, POSITIVITY and SUPERLATIVENESS play a less prominent role. We have found limited overlap with Caple *et al.* (2020), namely that NEGATIVITY is important in our study, while it has less importance in the Chinese data of their study. ELITENESS is equally important in both studies, while POSITIVITY and SUPERLATIVENESS are less often found in our study but prominent in theirs. Overall, our data confirm the importance of NEGATIVITY in the analysed news media. Thus, the analysis reveals unequivocal evidence that the media emphasised the extent of violence,

making terrorism appear the norm. It also suggests that ISIS was aligned with Islam and Muslims to create a negative perception of Islam in the world.

6. CONCLUSION

The present study aimed to contribute to the current knowledge of language and media discourse. The originality of this article can be summarised as follows. First, this is the first study that has investigated bilingual media reports, combining Arab media (*Al Jazeera* and *Al Arabiya* Arabic, and *Al Jazeera* and *Al Arabiya* English) and Western media (BBC and CNN in English) with DNVA analysis. The second significant contribution concerns the selection of the Libyan revolution outcomes. The focus was on the complex conflict that became an intricate, multi-layered civil war, with multiple sets of ideologies playing out in the background, which transformed into terrorism in the representation of the media. Other contributions of this study concern the methodological innovation. We not only applied a collocation analysis to retrieve the story contents within the studied media, but we also integrated the method of corpus techniques with DNVA. The former showed that the different media sets follow similar directions when reporting on the events and the latter explored news values constructed around the Libyan situation after the 2011 revolution with the analysis of keywords, collocations and concordances to provide significant results that emphasised NEGATIVITY and ELITENESS. Thus, it can be clearly stated that both Arab and Western media persistently reported the clashes between different Libyan factions during the years under investigation, as the Libyan situation worsened and turned into terrorism.

The media analysed here, with their violent and brutal discourse, tends only to show that Libya is living on the brink of collapse. In this sense, the current study shows how the Libyan civil war also transformed into terrorism in the representation of the media at a slow pace in 2013, but that it dominated the findings in 2015.

To sum up, we believe that further research can be conducted to examine local networks to understand the role of the media in framing world events, both politically and socially, in different cultural contexts. Also, we believe that the comparison between the media reports of violence and the discourse of terrorist groups can shed light on the way the media uses a similar discourse of violence to these groups and thus helps to promote the image of instability in the countries in conflict.

REFERENCES

- Anthony, Laurence. 2018. *AntConc* (Version 3.3.4) Tokyo: Waseda University. <https://www.laurenceanthony.net/software>
- Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.
- Baker, Paul. 2012. Acceptable bias? Using corpus linguistics methods with critical discourse analysis. *Critical Discourse Studies* 9/3: 247–256.
- Baker, Paul, Costas Gabrielatos, Majid Khosravini, Michał Krzyżanowski, Tony McEnery and Ruth Wodak. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse and Society* 19/3: 273–306.
- Bednarek, Monika. 2016. Voices and values in the news: News media talk, news values and attribution. *Discourse, Context & Media* 11: 27–37.
- Bednarek, Monika and Helen Caple. 2012a. *News discourse*. London: Continuum
- Bednarek, Monika and Helen Caple. 2012b. ‘Value added’: Language, image and news values. *Discourse, Context & Media* 1/2–3: 103–113.
- Bednarek, Monika and Helen Caple. 2014. Why do news values matter? Towards a new methodological framework for analyzing news discourse in critical discourse analysis and beyond. *Discourse & Society* 25/2: 135–158.
- Bednarek, Monika and Helen Caple. 2017. *The Discourse of News Values: How News Organizations Create Newsworthiness*. Oxford: Oxford University Press.
- Bell, Allan. 1991. *The Language of News Media*. Oxford: Blackwell.
- Bergen, Peter. 2006. *The Osama bin Laden I Know: An Oral History of Al Qaeda’s Leader*. New York: Free Press.
- Bhardwaj, Maya. 2012. Development of conflict in Arab Spring Libya and Syria: From revolution to civil war. *The Washington University International Review* 1/1: 76–96.
- Caple, Helen and Monika Bednarek. 2017. ‘What is DNVA?’ Discursive News Values Analysis (DNVA). <http://www.newsvaluesanalysis.com/what-is-dnva/> (21 October, 2020.)
- Caple, Helen, Changpeng Huan and Monika Bednarek. 2020. *Multimodal News Analysis across Cultures*. Cambridge: Cambridge University Press.
- Dahl, Trine and Kjersti Fløttum. 2017. Verbal-visual harmony or dissonance? A news values analysis of multimodal news texts on climate change. *Discourse, Context & Media* 20: 124–131.
- Doucet, Lyse. 2018. Syria & the CNN effect: What role does the media play in policy-making? *Daedalus* 147/1: 141–157.
- Fahmy, Shahira S. and Mohammed Al Emad. 2011. *Al-Jazeera vs Al-Jazeera: A comparison of the network’s English and Arabic online coverage of the US/Al Qaeda conflict*. *The International Communication Gazette* 73/3: 216–232.
- Fairclough, Norman. 2001. *Language and Power* (second edition). London: Longman.
- Firth, John. R. 1957. *Papers in Linguistics, 1934–1951*. Oxford: Oxford University Press.
- Fowler, Roger. G. 1991. *Language in the News: Discourse and Ideology in the Press*. London: Routledge.
- Freake, Rachelle, Guillaume Gentil and Jaffer Sheyholislami. 2010. A bilingual corpus-assisted discourse study of the construction of nationhood and belonging in Quebec. *Discourse & Society* 22/1: 21–47.
- Fuster Márquez, Miguel, and Carmen Gregori Signes. 2019. La construcción discursiva del turismo en la prensa española (verano de 2017). *Discurso y Sociedad* 13/2: 195–224.

- Garzone, Giuliana and Francesca Santulli. 2004. What can corpus linguistics do for critical discourse analysis? In Alan Partington, John Morley and Louann Haarman eds. *Corpora and Discourse*. Bern: Peter Lang, 351–368.
- Haider, Ahmad. S. 2016. A corpus-assisted critical discourse analysis of the representation of Qaddafi in media: Evidence from Asharq Al-Awsat and Al-Khaleej newspaper. *International Journal of Linguistics and Communication* 4/2: 11–29.
- Haigh, Michel M. and Michael Bruce. 2017. A comparison of the visual and story frames Al Jazeera English and CNN employed during the 2011 Egyptian Revolution. *International Communication Gazette* 79/4: 419–433.
- Hamdy, Naila and Ehab H. Gomaa. 2012. Framing the Egyptian uprising in Arabic language newspapers and social media. *Journal of Communication* 62/2: 195–211.
- Harcup, Tony and Deirdre O'Neill. 2001. What is news? Galtung and Ruge revisited. *Journalism Studies* 2/2: 261–280.
- Home Office. 2016. *Proscribed Terrorist Groups or Organisations*. <https://www.gov.uk/government/publications/proscribed-terror-groups-or-organisations--2> (23 June, 2020.)
- Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
- Irshaid, Faisal. 2014. Profile: Libya's Ansar Al-Sharia. *BBC News*. <http://www.bbc.com/news/world-africa-27732589> (23 June, 2020.)
- Kitano, Linus. 2019. *Constructing Allies Versus Non-allies in News Discourse: A Discursive News Values Analysis of US Media Reporting on Two Territorial Disputes*. Stockholm University MA Dissertation.
- Krishnamurthy, Ramesh and Wolfgang Teubert. 2007. General introduction. In Ramesh Krishnamurthy and Wolfgang Teubert eds. *Corpus Linguistics: Critical Concepts in Linguistics*. London: Routledge, 1–37.
- Liebes, Tamar and Anat First. 2004. Framing the Palestinian-Israeli conflict. In Norris Pippa, Montague Kern and Marion Just eds. *Framing Terrorism: The News Media, the Government and the Public*. London: Routledge, 59–74.
- Makki, Mohammad. 2019. Discursive news values analysis of Iranian crime news reports: Perspectives from the culture. *Discourse & Communication* 13/4: 437–460.
- Makki, Mohammad. 2020. The role of 'culture' in the construction of news values: A discourse analysis of Iranian hard news reports. *Journal of Multicultural Discourses* 15/3: 308–324.
- Maklad, Tawfik. S. 2019. Linguistic construction of news values in American news media coverage of a hate crime: A corpus-based discursive analysis. *مجلة البحث العلمي فى الآداب*: العدد العشرون الجزء الرابع (5), 637–715.
- Marsden, Lee and Heather Savigny. 2010. Introduction: Media, religion and conflict. In Lee Marsden and Heather Savigny eds. *Media, Religion and Conflict*. London: Routledge, 1–15.
- Mullet, Dianna R. 2018. A general critical discourse analysis framework for educational research. *Journal of Advanced Academics* 29/2: 116–142.
- Partington, Alan. 2006. Metaphors, motifs and similes across discourse types: Corpus-Assisted Discourse Studies (CADS) at work. In Anatol Stefanowitsch and Stefan Th. Gries eds. *Corpus-based Approaches to Metaphor and Metonymy*. Berlin: Mouton de Gruyter, 267–304.
- Peace and Security Council. 2011. *Communique of the 265th Meeting of the Peace and Security Council*. <https://www.peaceau.org/uploads/communique-libya-eng.pdf> (5 March, 2020.)

- Potts, Amanda, Monika Bednarek and Helen Caple. 2015. How can computer-based methods help researchers to investigate news values in large datasets? A corpus linguistic study of the construction of newsworthiness in the reporting on Hurricane Katrina. *Discourse & Communication* 9/2: 149–172.
- Richardson, John E. 2006. *Analysing Newspapers: An Approach from Critical Discourse Analysis*. London: Bloomsbury.
- Romero-Trillo, Jesús. 2011. The representation of liminality conflicts in the media. *Journal of Multicultural Discourses* 6/2: 143–158.
- Romero-Trillo, Jesús and Caroline Cheshire. 2014. The construction and disarticulation of national identities through language vis-a-vis the Scottish Referendum of Independence. *Lodz Papers in Pragmatics* 10/1: 41–66.
- Romero-Trillo, Jesús and Safa Attia. 2016. Framing the ideological outcomes of the Tunisian Revolution through the eyes of the Arab and Western media. *Lodz Papers in Pragmatics* 12/2: 177–213.
- Schulz, Winfried Friedrich. 1982. News structure and people's awareness of political events. *International Communication Gazette* 30/3: 139–153.
- Smits, Rosan, Floor Janssen, Ivan Briscoe and Terri Beswick. 2013. *Revolution and its Discontents: State, Factions and Violence in the New Libya*. The Hague: Netherlands Institute of International Relations Clingendael.
- Stubbs, Michael. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.
- Stubbs, Michael. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.
- Taylor, Charlotte. 2013. Searching for similarity using corpus-assisted discourse studies. *Corpora* 8/1: 81–113.
- Törnberg, Anton and Petter Törnberg. 2016. Muslims in social media discourse: Combining topic modeling and critical discourse analysis. *Discourse, Context & Media* 13/B: 132–142.
- Van Dijk, Teun A. 1997. What is political discourse analysis? *Belgian Journal of Linguistics* 11/1: 11–52.
- Wodak, Ruth. 2001. The discourse-historical approach. *Methods of Critical Discourse Analysis* 1: 63–94.
- Wodak, Ruth. 2004. Critical discourse analysis. In Clive Seale, Giampietro Gobo, Jaber F. Gubrium and David Silverman eds. *Qualitative Research Practice*. London: SAGE, 197–213.
- Yilmaz, Mesut and Oktay Sinanoglu. 2014. The effect of dominant ideology on media: The Syria case. *The European Journal of Social & Behavioural Sciences* 10/3: 1527–1541.

Corresponding author

Safa Attia

Autonomous University of Madrid

e-mail: safa.attia@estudiante.uam.es

received: August 2021

accepted: January 2022

The FGLOCTweet Corpus: An English tweet-based corpus for fine-grained location-detection tasks

Nicolás José Fernández-Martínez
Catholic University of Murcia / Spain

Abstract – Location detection in social-media microtexts is an important natural language processing task for emergency-based contexts where locative references are identified in text data. Spatial information obtained from texts is essential to understand where an incident happened, where people are in need of help and/or which areas have been affected. This information contributes to raising emergency situation awareness, which is then passed on to emergency responders and competent authorities to act as quickly as possible. Annotated text data are necessary for building and evaluating location-detection systems. The problem is that available corpora of tweets for location-detection tasks are either lacking or, at best, annotated with coarse-grained location types (e.g. cities, towns, countries, some buildings, etc.). To bridge this gap, we present our semi-automatically annotated corpus, the *Fine-Grained LOcation Tweet Corpus* (FGLOCTweet Corpus), an English tweet-based corpus for fine-grained location-detection tasks, including fine-grained locative references (i.e. geopolitical entities, natural landforms, points of interest and traffic ways) together with their surrounding locative markers (i.e. direction, distance, movement or time). It includes annotated tweet data for training and evaluation purposes, which can be used to advance research in location detection, as well as in the study of the linguistic representation of place or of the microtext genre of social media.

Keywords – location detection; locative references; fine-grained locations; tweets; corpus for training and evaluating models

1. INTRODUCTION

Location detection is an important task in Natural Language Processing (NLP) whereby locative references mentioned in texts are identified and extracted for a variety of practical purposes (Middleton *et al.* 2018; Purves *et al.* 2018). This task has recently been applied to microtext genres such as tweets which, due to their brief and informal nature, contain many non-standard forms that challenge the performance of current NLP systems which are typically trained on more formal genres such as news (Baldwin *et al.* 2013; Eisenstein 2013). Hence, there is an increasing need to focus on building and using corpora based on social media microtexts.



Location detection from social media microtexts has wide-ranging practical applications: from natural or human-made disaster detection and tracking in floods, earthquakes, storms, civil unrest, war, crime, etc. (Vieweg *et al.* 2010; Crooks *et al.* 2013; Imran *et al.* 2014; Jongman *et al.* 2015; Martínez-Rojas *et al.* 2018; Siriaraya *et al.* 2019; Zhang *et al.* 2019), health surveillance and disease tracking (Eke 2011; Dredze *et al.* 2013), for example, the COVID-19 pandemic (Singh *et al.* 2020), to marketing and advertising purposes (Mourad *et al.* 2019), or traffic incident detection, road traffic control and/or traffic congestion (Ahmed *et al.* 2019; Das and Purves 2019; Gonzalez-Paule *et al.* 2019; Khodabandeh-Shahraki *et al.* 2019). In this regard, the extraction of fine-grained geospatial information from social media microtexts plays a key role in intelligent systems for crisis management services to raise emergency situation awareness from crisis-related events where the location dimension proves essential to understand their impact: where an incident happened, where people are in need of help and/or which areas have been affected (Vieweg *et al.* 2010; Crooks *et al.* 2013; Imran *et al.* 2014). Such information could potentially help save lives and/or prevent further damage to environmental or urban areas in emergency- and crisis-related contexts.

Corpus building in this area helps train location-detection systems in supervised probabilistic-based models, typically based on machine learning or deep learning, or develop rule-based systems and assess their performance, with a view to replicating their performance in real-life contexts. The problem is that (i) most tweet corpora are not available for public use, impeding any replication or future development, and (ii) that corpus development in this area has extensively focused on annotating coarse-grained location types such as geopolitical entities (e.g. countries, cities or towns), leaving aside many other toponymic entities that are equally, if not more, important in crisis-related scenarios, such as points of interests, natural landforms and traffic ways. Also, information related to distance, direction or time that may accompany such entities is not tagged, losing again the granularity needed for emergency-based services.

To address these issues, we present the *Fine-Grained LOcation Tweet Corpus* (FGLOCTweet Corpus), which has been semi-automatically built using our linguistically aware location-detection system LOcative Reference Extractor (LORE) for its processing and annotation (Fernández-Martínez and Perrián-Pascual 2021a), including the anonymization of users' references, and supervised error revision. The corpus integrates English tweets with annotated coarse- and fine-grained locative references from real-life

situations for the development and evaluation of location-detection systems with an interest in a greater diversity, variety and semantic granularity of the location types. We may release the corpus upon request¹ for its use in location-detection research development or for linguistic inquiry of the microtext genre and the representation of spatial knowledge in English.

The present article is structured as follows. First, we briefly examine related work in tweet location detection paying special attention to the corpora used, their characteristics and their availability. Then, we provide the methodological steps in building and annotating our corpus. Finally, we discuss the practical uses and applications for research practitioners and conclude with some future research directions.

2. BRIEF OVERVIEW OF THE LITERATURE

We provide here, in chronological order, some of the major contributions in corpus building for tweet-based location detection tasks. Most authors have built their own corpus containing thousands of tweets focusing on geopolitical entities (e.g. cities, towns and countries), and have typically restricted themselves to specific areas or crisis-related events (Inkpen *et al.* 2017; de Bruijn *et al.* 2018). However, most of these self-compiled corpora are unavailable for public use, and they contain, most of the times, coarse-grained locative references only. Other problems relate to the use of different corpus annotation standards, which aggravates the reutilization of such resources.

Inkpen *et al.* (2017) built, for their probabilistic-based location-detection system, a corpus of 6,000 semi-automatically annotated tweets containing 4,369 mentions of coarse-grained locations (i.e. cities, provinces and states) from the US and Canada.² The building process consisted of two phases: first, a simple matching step was performed using the *GeoNames* database (Ahlers 2013) to match names of locations from the US and Canada together with their location type and, then, a manual filtering process by expert annotators was conducted to revise errors or include other missed entities. Their corpus missed richer locative reference types such as points of interests, streets or

¹ According to *Twitter's* Privacy and Developer policies, “[...] all developers may provide up to 50,000 public Tweets Objects and/or User Objects to each person who uses your service on a daily basis if this is done via non-automated means (e.g., download of spreadsheets or PDFs).” (Developer Policy – *Twitter* Developers 2021). This means that we can only share these tweets upon users’ requests for non-for-profit purposes.

² Available at <https://github.com/rex911/locdet> (5 July, 2021).

highways. Likewise, other important locative markers were ignored (e.g. distance markers such as *n kilometres away from X*), and it only focused on US and Canada entities, leaving aside many other geographic areas of the world.

De Bruijn *et al.* (2018) compiled 2,785 flood-related tweets, manually tagging geopolitical entities such as countries, cities, towns and villages from different parts of the world up to a number of 2,079 locative references mentioned in those tweets, by using a matching algorithm and the *GeoNames* database.³ Since only geopolitical entities were labeled, the corpus lacks a great deal of fine-grained locative references and potential locative markers.

The most famous available corpus for location-detection purposes is the *GeoCorpora*, built by Wallgrün *et al.* (2018).⁴ *GeoCorpora* contains 6,711 tweets of a variety of crisis-related events using keywords as diverse as *floods*, *riots*, *tornados*, *flu*, *violence*, etc. with their correspondingly mentioned locative references, a unique ID and geographic coordinates obtained from the *GeoNames* database, whenever available. Geographers were used to tag and revise the annotation of place names. Since only tweet IDs are provided to retrieve the tweets, it is possible that many may have been deleted by now. The locative types considered in the annotation of the corpus were mostly towns, cities, states and countries, as well as some natural landforms and a few traffic ways (e.g. street names and addresses). Sometimes, location-indicative nouns (e.g. lake, hill, county or state) were tagged as part of locative references. However, the corpus lacks a great deal of location types, and locative markers are not considered.

Hu and Wang (2020) obtained, preprocessed and annotated 1,000 tweets out of a very large corpus of 7,041,866 tweets collected in the event of the Hurricane Harvey in the US in 2017.⁵ They performed a study of the location types mentioned in those tweets, differentiating the following: addresses, street names, highways, exit of highways, roads, natural landforms, buildings and geopolitical entities of different types. They also assessed general-domain entity recognizers and found that they fail at detecting traffic ways tremendously. To this date, this is the only released corpus providing a number of easily accessible annotated fine-grained locative references. However, its focus is on a

³ The code is available at <https://github.com/jensdebruijn/TAGGS> (5 July, 2021.), but the corpus is not publicly available.

⁴ Available at <https://github.com/geovista/GeoCorpora> (10 September, 2021.).

⁵ Available at <https://github.com/geoai-lab/HowDoPeopleDescribeLocations> (10 September, 2021.).

particular event in a specific area, thus limiting its application to any other crisis-related event in other parts of the world.

Recent research highlights that the task of location detection in social microtexts is not a solved task (Wang and Hu 2019). Specifically, it has been mentioned that there is an ever-growing need for detecting fine-grained locations (street names or the names of parks and monuments), as well as developing corpora of social media microtexts for training and evaluating models with fine-grained locative references (Gritta *et al.* 2018). Considering such limitations in the state of the art, the FGLOCTweet Corpus is intended to address this gap by providing a great number of tweets with annotated fine-grained locative references from a variety of incidents and crisis-related events from all over the world.

3. METHODOLOGY

Corpus building involved a series of steps that will be discussed in further detail later in this section. These include corpus compilation, corpus preprocessing and corpus annotation (Rayson 2014: 33). All these stages were performed with semiautomatic techniques (regular expressions, automatic tagging, and manual revision) that greatly facilitated the construction of the corpus. Also, the corpus was built with methodological corpus-based principles in mind in each of the different steps, including size, representativeness and balance, as well as practicality (Reppen 2010). To be more specific, the FGLOCTweet Corpus was built with the aim of capturing as wide a variety of locative references as possible from as many different real-life incidents from as many places in the world. A total of 9,405 tweets with their corresponding tagged tokens seemed to be the sweet spot for locative-detection tasks, since the size of such a specialized corpus does not need to be particularly large but surely needs to be sufficiently representative, in accordance with the corpus size found in the literature of location detection.

In the case of our corpus, which was built for NLP applications, it was also important to consider consistency in the annotation of data or, in other words, that a set of guidelines was followed for the correct training of our model and also for a correct evaluation of the model (Zinsmeister *et al.* 2009: 764). This meant using the same set of part-of-speech (POS) tags. As for the annotated locative references, this implied that these

adhere to i) morphological, structural and semantic criteria involving a definition of what a locative reference is, and ii) the degree of geospatial granularity needed for real-life emergency-based applications. We define a locative reference as a proper noun that designates a geographically locatable spatial point, morphologically realized with full names, abbreviations, acronyms or a given combination of all of them. Structurally speaking, they are either simple or complex, depending on the number of lexical and/or phrasal elements accompanying the proper noun(s). This is illustrated in Figure 1, where an asterisk is used to mark optionality and double asterisk refers to the optional presence of locative markers either at the beginning or at the end of the locative reference.

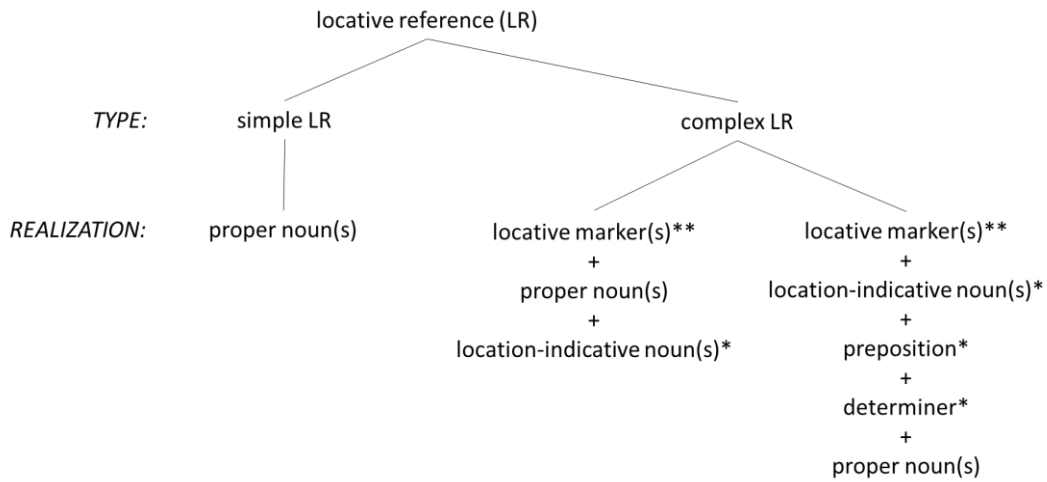


Figure 1: Phrasal structure of locative references

Examples of locative references are *Morocco*, *New York*, *south of Madrid*, *1km SW of Lake Henshaw*, *1h away from London*, *25min out of Melbourne*, *Mount Crawford*, *Bassmaya Project Power Plant station*, *province of Ontario*, *Jamia University*, *Dyckman Street Station*, *St Albans Park*, *Fox Valley Animal Referral Center*, *I 95 NB* and *George Washington bridge EB*.

In terms of semantics, a taxonomy was devised capturing the richness and variety of locative references, where abbreviations were also taken into account:

- Geopolitical entities: *country*, *state*, *region*, *province*, *city*, *town*, *kingdom*, *villa*, etc.
- Natural landforms: *mountain*, *mount*, *ridge*, *volcano*, *valley*, *lake*, *river*, *shore*, *beach*, *park*, *canyon*, etc.
- Points of Interest (POIs): *building*, *museum*, *school*, *station*, *stadium*, *garden*, *café*, *tavern*, *hospital*, *court*, *theater*, *residence*, *zoo*, *casino*, *square*, etc.

- Traffic ways (addresses, roads, highways): *street, st, boulevard, blvd, avenue, av, alley, road, rd, highway, hwy, freeway, fwy, turnpike, tpk, I(-)n, M(-)n* (where *n* represents a given number), etc.

Metonymic instances are a well-known issue in the literature when these represent the people of a place (e.g. *US officials*), organizations (e.g. *New Orleans Police Department*), government units (e.g. *London Councils*) or events (e.g. *New Zealand mass shooting*) (Liu *et al.* 2014; Gritta *et al.* 2018). They are borderline cases of locative references, and semantic boundaries are hard to establish (Wallgrün *et al.* 2018). A solution for this issue consisted in examining every ambiguous instance and determining, on the basis of the linguistic context, whether some degree of locative meaning was found in those references.

Given the importance of surrounding locative markers (e.g. *south of, northwest, 25km away from, 20 mins out of*, etc.), which contain more detailed information about the locational focus of a given incident, these were also annotated, following this taxonomy:

- Distance marker: *4 Kms from Narok Town, 5miles from Dublin*, etc.
- Directional markers: *East Coast of Honshu, east of Exit 55, 20 km NW of Durrës*, etc.
- Movement markers: *southbound I-91, northbound J19, eb J19*, etc.
- Temporal markers: *1h away from London, 25min out of Melbourne*, etc.

3.1. Corpus compilation

The first stage involves decisions about text collection and corpus design. For collecting the tweets, which are the microtexts used in our corpus, the *FireAnt* tool was used (Anthony and Hardaker 2017) to obtain machine-readable tweet data, that is, JavaScript Object Annotation (JSON). The raw tweet data were collected on different dates after a keyword-based search using seven keywords related to crisis and emergency-related events which were *earthquake, flood, car accident, bombing attack, shooting attack, terrorist attack* and *incident*, so that tweets mentioning issues of different nature were extracted. The dates of extraction of the tweets were the following: 17 November 2019, 30 November 2019, 1 December 2019 and 2, 5 and 9 January 2020. In the corpus design substep, we also tackled key considerations such as what file formats were to be used and

what type of information would be included therein. The *FireAnt* tool provided not only the tweet text, but also the metadata associated with it. All those metadata were discarded and only tweet texts were saved in a .txt file before the preprocessing stage.

3.2. Corpus preprocessing

While the great majority of tweet texts contained one of the crisis-related keywords mentioned above, it was the case that some tweets were repeated on multiple occasions in retweets, split into different lines or empty. Our aim in this step was to obtain a representative corpus of unique tweets. The first preprocessing stage thus involved the following steps:

- i) grouping multi-line tweets into a single line where each line represented one tweet by means of a regular expression that takes into account line breaks,
- ii) removing retweets by means a regular expression that finds retweets and discards them, and
- iii) removing duplicates and very similar tweets through a fuzzy matching algorithm (i.e. cosine distance similarity), which takes into account different combinations of characters and words in two strings to determine their degree of similarity.

Even though sensitive data about the tweets were removed by retaining the tweet text only, the text still contained sensitive information in the form of user mentions and URLs, which were dealt with in a second preprocessing stage. In this second preprocessing stage, non-standard linguistic features were kept in the tweets, too, since a key aspect in tweet location detection is to be able to overcome the challenge posed by non-standard uses of language. The main steps followed were the following:

- i) Replacing user mentions and URLs by the tokens *user* and *URL*, respectively.
- ii) Removing emojis and other special characters and leave punctuation marks and other commonly used characters (/ , @ , | ...).
- iii) Removing extra white spaces.
- iv) Segmenting words contained in hashtags.

After both preprocessing stages, the resulting tweets were as unique as possible, clean and privacy-friendly. The released corpus contains the preprocessed tweets as such.

3.2. Corpus annotation

In the annotation stage, the corpus content was converted into a token-based tabular format with feature columns, separated by tabs, representing the following features: token, POS tag, presence in a *GeoNames*-based place-name dataset, presence in *WordNet*-based location-indicative noun dataset and part of a locative marker. In the last column, the class or label was tagged, following a Beginning-Medium-End-Single-Outside (BMESO) scheme, similar to others in Named Entity Recognition (NER), such as Beginning-Inside-Outside (BIO) (Jurafsky and Martin 2021). In other words, for multi-word locative references, the following labels were used: B_LOCATION, M_LOCATION and E_LOCATION. In the case of one-word locative references, S_LOCATION was used and, when a token or series of tokens are not locative references, O was used.

First, for preparing the annotated corpus, automatic tokenization and POS tagging were automatically applied, using the Stanford tokenizer and POS tagger functionalities (Toutanova and Manning 2000), and each tokenized tweet was delimited by a newline. The POS tags followed the *Penn Treebank* standard (Santorini 1990). Then, if tokens or a series of them were found in a place-name dataset obtained from *GeoNames* or in the location-indicative noun dataset obtained from *WordNet* (Vossen 1998) or were part of a locative marker, they were also automatically annotated with Boolean values: 0 if not present, 1 if present. This automatic tagging process was performed with the linguistic modules of LORE (Fernández-Martínez and Periñán-Pascual 2021a), which also, at last, detected and tagged the locative references found in the tweets. Table 1 presents an example of the token-based tabular format of the annotated corpus.

Token	POS tag	Place-name dataset	Location-indicative noun dataset	Locative marker	Label
Two	CD	0	0	0	O
vehicle	NN	0	0	0	O
incident	NN	0	0	0	O
,	,	0	0	0	O
48	CD	0	0	0	B_LOCATION
St	NNP	0	1	0	E_LOCATION
and	CC	0	0	0	O
32	CD	0	0	0	B_LOCATION
Ave	NN	1	1	0	M_LOCATION
NE	NNS	1	0	1	E_LOCATION

Table 1: Token-based tabular format in the FGLOCTweet Corpus

Since the accuracy of LORE is not perfect for detecting all and only locative references (precision score of 0.81 and recall score of 0.81 in Fernández-Martínez and Perinián-Pascual 2021a; precision score of 0.73 and recall score of 0.79 in Fernández-Martínez and Perinián-Pascual 2021b), the tags were manually revised for errors, such as missed locative references or wrongly assigned locative references, on the basis of the guidelines of the location types targeted by LORE. This was done using a common notepad editor tool (Notepad++). The POS tags were not revised since i) automatic POS tagging tools achieve a very high degree of accuracy (Manning 2011) and ii) feature noise is not a problem *per se* as long as other features can contribute in the learning process of the probabilistic-based model (Zhu and Wu 2004). POS tagging is a common component in NLP tasks and a typical feature, alongside tokenization, used in existing NER, since POS tags provide a strong linguistic cue for predicting the presence of named entities (Jurafsky and Martin 2021), especially because named entities are proper nouns. Particularly, in the case of locative references, these can be predicted by the presence of spatial prepositions (Hoang and Mothe 2018). However, automatic POS tagging might suffer from performance losses especially in the case of noisy text data (e.g. abbreviations, misspellings, ellipsis, etc.).

The other three features, presence in the place-name dataset obtained from *GeoNames*, presence in the location-indicative noun dataset retrieved from *WordNet* and being part of a locative marker, might also be prone to noise if a token or series of tokens are not found in these datasets. However, since different features are correlated, this noise might have a negligible impact in the training and evaluation phases of a location-detection system.⁶

Table 2 presents the distribution of locative references in terms of n-gram size in the corpus, whereas Table 3 provides statistical data about the number of locative references, number of tweets containing locative references, the average of locative references per tweet containing locative references and the average of locative references per tweet.

⁶ In fact, it is known that in probabilistic-based models some degree of noise in a dataset can even be beneficial to avoid overfitting, that is, the memorization of the training dataset at the cost of performance degradation with new, unseen instances of data (Zur *et al.* 2009).

Number of unigrams (e.g. <i>Florida</i>)	3,256
Number of bigrams (e.g. <i>Grand Canyon</i>)	1,707
Number of trigrams (e.g. <i>St Albans Park</i>)	501
Number of n-grams where $n \geq 4$ (e.g. <i>Fox Valley Animal Referral Center</i>)	304
Total	5,768

Table 2: Distribution of locative references in terms of n-gram size in the corpus

Number of locative references	5,768
Number of tweets with locative references	3,416
Average of locative references per locative-rich tweet	1.69
Average of locative references per tweet	0.61

Table 3: Statistics about the number and average of locative references in the corpus

4. DISCUSSION: APPLICATIONS OF THE CORPUS, LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

The resulting corpus can then be split into two subcorpora: training and evaluation corpora using, roughly, an 85/15 split, which is the rule of thumb in the machine learning literature (Guyon 1997). The training corpus can be used to train a supervised probabilistic-based model for location detection, whereas the evaluation corpus can be used as a gold standard against which the output of a location-detection system can be tested for the evaluation of its accuracy (Pustejovsky and Stubbs 2013).

The main use of the FGLOCTweet Corpus is to build and assess location-detection models, either rule-based or probabilistic-based, for the task of identifying fine-grained locative references in crisis-related events from all over the world. Fine-grained detection of locative references is indeed a key aspect of accurate and useful location-detection systems which could potentially be used to save lives or prevent further damage to environmental or urban areas in crisis-related events by providing emergency responders with the location of a given incident. The corpus could also be used as a benchmarking dataset to compare the performance of different location-detection models, including named entity recognizers, too. Beyond that, linguists may find this corpus useful for approaching the conceptualization, expression and description of place in English during crisis-related events or even a general exploration of language use in microtexts dealing with crisis-related events.

In past research (Fernández-Martínez and Perinán-Pascual 2021a, 2021b), LORE, a rule-based model, and its probabilistic-based counterpart, neural LORE (nLORE), were built and assessed using this corpus, outperforming general-domain entity recognizers in benchmarking tests involving accuracy (i.e. precision and recall) and speed. A key

difference in the implementation of both models lies in how they make use of corpus data: the probabilistic-based model nLORE needed training data before the evaluation stage with the evaluation corpus, whereas the rule-based model LORE did not. This means that probabilistic-based models consume a lot of computational resources and time, whereas rule-based models tend to be much more efficient and quicker (Chiticariu *et al.* 2013). In this regard, LORE performed ten times faster than nLORE: it extracted locative references from around 7,000 tweets in a matter of 12 seconds as opposed to nLORE, for which it took almost two minutes. However, nLORE had slightly better accuracy than LORE in terms of precision (0.85 vs. 0.73), but lower recall (0.74 vs. 0.79).

As for the limitations, we would like to emphasize that, even though the Stanford POS tagger may achieve a very high accuracy of 97 percent (Manning 2011), its accuracy might have been somewhat lower with the tweets, introducing some corpus noise in the POS tags feature. A rule-based model that is assessed on our corpus, if developed with lexicogrammatical rules taking into account grammatical categories, might be misled by wrong POS tags and extract wrong items or miss potential locative references, if it does not rely on the other corpus features as well. The probabilistic-based model would not, however, be impeded by this providing that it makes use of the different features at the same time, and even in that case, some degree of corpus noise, as stated above, might be beneficial to avoid overfitting in the training phase of the probabilistic-based model.

Further research lines could be pursued, especially in a time where novel NLP approaches employing transformers like BERT (Devlin *et al.* 2018) show promising results, which could be fine-tuned using our corpus. Also, LORE could be employed to automatically aggregate new annotated data to the corpus in an unsupervised fashion, thus enriching the number and variety of locative references, though at the cost of introducing corpus noise. In such a scenario, it can be insightful to analyze whether corpus noise in larger sizes of annotated corpus data could be detrimental to the performance of a model trained on and assessed with these data. Also, transfer learning techniques together with this unsupervised aggregate of new data could be used to create a multilingual corpus for multilingual fine-grained location-detection tasks, including and mixing, for instance, annotated Spanish and French tweets.

5. CONCLUSION

Location detection in social media is still an unsolved task in NLP. Since there is a growing need to automatically obtain actionable information from social media in emergency-based contexts where granularity and time play an essential role to understand the *where* of an incident, the development of fine-grained annotated corpus data becomes of utmost importance, especially to train and assess location-detection models. Despite that, available corpora are lacking up to this date; annotation standards are different, and many location types are, at best, poorly addressed in the literature or, at worst, neglected. Besides, phrasal structures indicating distance (e.g. *n kms away from X*), direction (e.g. *south of X*), movement (e.g. *eastbound*) and time (e.g. *n mins out of X*), which may take part in locative references, are not annotated, thus missing very detailed geospatial information which could be highly important in crisis-related events. To address this gap, the FGLOCTweet Corpus is presented, with the aim of providing an English tweet-based corpus for fine-grained location-detection tasks to advance research in the NLP and linguistic communities.

REFERENCES

- Ahlers, Dirk. 2013. Assessment of the accuracy of GeoNames gazetteer data. In Chris Jones and Ross Purves eds. *Proceedings of the 7th Workshop on Geographic Information Retrieval - GIR '13*. New York: Association for Computing Machinery, 74–81.
- Ahmed, Mohammed F., Lelitha Vanajakshi and Ramasubramanian Suriyanarayanan. 2019. Real-time traffic congestion information from tweets using supervised and unsupervised machine learning techniques. *Transportation in Developing Economies* 5/2: Article 20. <https://link.springer.com/article/10.1007/s40890-019-0088-2> (10 September, 2021.)
- Anthony, Laurence and Claire Hardaker. 2017. *FireAnt* (Version 1.1.4). Tokyo, Japan: Waseda University. <https://www.laurenceanthony.net/software> (10 September, 2021.)
- Baldwin, Timothy, Paul Cook, Marco Lui, Andrew MacKinlay and Li Wang. 2013. How noisy social media text, how diffrent social media sources? In Ruslan Mitkov and Jong C. Park eds. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Nagoya, Japan: Asian Federation of Natural Language Processing, 356–364. <http://www.aclweb.org/anthology/I13-1041> (10 September, 2021.)
- Chiticariu, Laura, Yunyao Li and Frederick R. Reiss. 2013. Rule-based information extraction is dead! Long live rule-based information extraction systems! In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu and Steven Bethard eds. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. New York: Association for Computational Linguistics, 827–

832.

- Crooks, Andrew, Arie Croitoru, Anthony Stefanidis and Jacek Radzikowski. 2013. #Earthquake: Twitter as a distributed sensor system. *Transactions in GIS* 17/1: 124–147.
- Das, Raul D. and Ross S. Purves. 2019. Exploring the potential of Twitter to understand traffic events and their locations in greater Mumbai, India. *IEEE Transactions on Intelligent Transportation Systems* 21/12: 1–10.
- De Bruijn, Jens A., Hans de Moel, Brenden Jongman, Jurgen Wagemaker and Jeroen C. Aerts. 2018. TAGGS: Grouping tweets to improve global geoparsing for disaster response. *Journal of Geovisualization and Spatial Analysis* 2/2: 1–14.
- Developer Policy – Twitter Developers. 2021. Twitter developer platform. <https://developer.twitter.com/en/developer-terms/policy> (5 December, 2021.)
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv* <http://arxiv.org/abs/1810.04805> (10 September, 2021.)
- Dredze, Mark, Michael J. Paul, Shane Bergsma and Hieu Tran. 2013. Carmen: A twitter geolocation system with applications to public health. In Martin Michalowski, Wojtek Michalowski, Dymrna O’Sullivan, Szymon Wilk eds. *Expanding the Boundaries of Health Informatics Using Artificial Intelligence: Papers from the Association for the Advancement of Artificial Intelligence 2013 Workshop*. Palo Alto, California: Association for the Advancement of Artificial Intelligence, 20–24. <https://www.aaai.org/ocs/index.php/WS/AAAIW13/paper/view/7085>
- Eisenstein, Jacob. 2013. What to do about bad language on the internet. In Lucy Vanderwende, Hal Daumé III and Katrin Kirchhoff eds. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New York: Association for Computational Linguistics, 359–369. <https://aclanthology.org/N13-1037/>
- Eke, Paul I. 2011. Using social media for research and public health surveillance. *Journal of Dental Research* 90/9: 1045–1046.
- Fernández-Martínez, Nicolás José and Carlos Perinián-Pascual. 2021a. LORE: A model for the detection of fine-grained locative references in tweets. *Onomazein* 52: 195–225.
- Fernández-Martínez, Nicolás José and Carlos Perinián-Pascual. 2021b. nLORE: A linguistically rich deep-learning system for locative-reference extraction in tweets. In Engie Bashir and Mitja Luštrek eds. *Intelligent Environments 2021: Workshop Proceedings of the 1st International Workshop on Artificial Intelligence and Machine Learning for Emerging Topics (ALLEGET ’21)*. Amsterdam: IOS Press 243–254.
- Gonzalez-Paule, Jorge David, Yeran Sun and Yashar Moshfeghi. 2019. On fine-grained geolocalisation of tweets and real-time traffic incident detection. *Information Processing and Management* 56/3: 1–14.
- Gritta, Milan, Moahammad T. Pilehvar, Nut Limsopatham and Nigel Collier. 2018. What’s missing in geographical parsing? *Language Resources and Evaluation* 52/2: 603–623.
- Guyon, Isabelle. 1997. A scaling law for the validation-set training-set size ratio. Technical report. Berkeley, California: AT&T Bell Laboratories 1–11. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.1337&rep=rep1&type=pdf> (10 September, 2021.)
- Hoang, Thi B. N. and Josiane Mothe. 2018. Location extraction from tweets. *Information Processing and Management* 54/2: 129–144.

- Hu, Yingjie and Jimin Wang. 2020. How do people describe locations during a natural disaster: An analysis of tweets from Hurricane Harvey. In Krzysztof Janowicz and Judith A. Versteegen eds. *11th International Conference on Geographic Information Science (GIScience 2021)*. Dagstuhl, Germany: Dagstuhl Publishing Company, 6.1–6.16. <https://drops.dagstuhl.de/opus/volltexte/2020/13041/pdf/LIPIcs-GIScience-2021-I-6.pdf> (10 September, 2021.)
- Imran, Muhammad, Carlos Castillo, Fernando Diaz and Sarah Vieweg. 2014. Processing social media messages in mass emergency: Survey summary. *WWW'18: Companion Proceedings of the The Web Conference 2018*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 507–511. <https://dl.acm.org/doi/10.1145/3184558.3186242> (10 September, 2021.)
- Inkpen, Diana, Ji Liu, Atefeh Farzindar, Farzaneh Kazemi and Diman Ghazi. 2017. Location detection and disambiguation from twitter messages. *Journal of Intelligent Information Systems* 49/2: 237–253.
- Jongman, Brenden, Jurgen Wagemaker, Beatriz Romero and Erin de Perez. 2015. Early flood detection for rapid humanitarian response: Harnessing near real-time satellite and Twitter signals. *ISPRS International Journal of Geo-Information* 4/4: 2246–2266.
- Jurafsky, Daniel and James H. Martin. 2021. Sequence labeling for parts of speech and named entities. In Dan Jurafsky and James H. Martin eds. *Speech and Language Processing*: 1–27. <https://web.stanford.edu/~jurafsky/slp3/8.pdf> (10 September, 2021.)
- Khodabandeh-Shahraki, Zahra, Afsaneh Fatemi and Hadi Tabatabaee-Malazi. 2019. Evidential fine-grained event localization using Twitter. *Information Processing and Management* 56/6: Article 102045.
- Liu, Fei, Maria Vasardani and Timothy Baldwin. 2014. Automatic identification of locative expressions from social media text. In Dirk Ahlers ed. *LocWeb '14: Proceedings of the 4th International Workshop on Location and the Web*. New York: Association for Computing Machinery, 9–16.
- Manning, Christopher D. 2011. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In Alexander F. Gelbukh ed. *Computational Linguistics and Intelligent Text Processing*. Berlin: Springer Berlin Heidelberg, 171–189.
- Martínez-Rojas, María, María del Carmen Pardo-Ferreira and Juan Carlos Rubio-Romero. 2018. Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. *International Journal of Information Management* 43: 196–208.
- Middleton, Stuart E., Giorgos Kordopatis-Zilos, Symeon Papadopoulos and Yiannis Kompatsiaris. 2018. Location Extraction from Social Media. *ACM Transactions on Information Systems* 36/4: 1–27.
- Mourad, Ahmed, Falk Scholer, Walid Magdy and Mark Sanderson. 2019. A practical guide for the effective evaluation of Twitter user geolocation. *ACM Transactions on Social Computing* 2/3: 1–23.
- Purves, Ross S., Paul Clough, Christopher B. Jones, Mark H. Hall and Vanessa Murdock. 2018. Geographic information retrieval: Progress and challenges in spatial search of text. *Foundations and Trends in Information Retrieval* 12/2–3: 164–318.
- Pustejovsky, James and Amber Stubbs. 2013. *Natural Language Annotation for Machine Learning: A Guide to Corpus-building for Applications*. Sebastopol, California: O'Reilly Media, Inc.
- Rayson, Paul. 2014. Computational tools and methods for corpus compilation and

- analysis. In Douglas Biber and Randi Reppen eds. *The Cambridge Handbook of English Corpus Linguistic*. Cambridge: Cambridge University Press, 32–50.
- Reppen, Randi. 2010. Building a corpus. In Anne O’Keeffe and Michael McCarthy eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 31–37.
- Santorini, Beatrice. 1990. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project*. 3rd revision, 2nd printing. Department of Computer and Information Science, University of Pennsylvania: Technical Report MS-CIS-9047. https://repository.upenn.edu/cis_reports/570/ (10 September, 2021.)
- Singh, Lisa, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga and Yanchen Wang. 2020. A first look at COVID-19 information and misinformation sharing on Twitter [preprint 31 March 2020]. *ArXiv*. <http://arxiv.org/abs/2003.13907> (10 September, 2021.)
- Siriaraya, Panote, Yihong Zhang, Yuanyuan Wang, Yukiko Kawai, Mohit Mittal, Péter Jeszenszky and Adam Jatowt. 2019. Witnessing crime through tweets. In Farnoush Banaei-Kashani, Goce Trajcevski, Ralf Hartmut Güting, Lars Kulik and Shawn Newsam eds. *SIGSPATIAL ’19: Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. New York: Association for Computing Machinery, 568–571.
- Toutanova, Kristina and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In Hinrich Schütze and Keh-Yih Su eds. *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. New York: Association for Computational Linguistics, 63–70.
- Vieweg, Sarah, Amanda L. Hughes, Kate Starbird and Leysia Palen. 2010. Microblogging during two natural hazards events. In Elizabeth Mynatt ed. *CHI ’10 Proceedings of the 28th International Conference on Human Factors in Computing Systems*. New York: Association for Computing Machinery, 1079–1088.
- Vossen, Piek. 1998. Introduction to EuroWordNet. *Computers and the Humanities* 32/2–3: 73–89.
- Wallgrün, Jan Oliver, Morteza Karimzadeh, Alan M. MacEachren and Scott Pezanowski. 2018. GeoCorpora: Building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science* 32/1: 1–29.
- Wang, Jimin and Yingjie Hu. 2019. Are we there yet? Evaluating state-of-the-art neural network based geoparsers using EUPEG as a benchmarking platform. In Bruno Martins ed. *GeoHumanities ’19 Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities*. New York: Association for Computing Machinery, Article 2, 1–6.
- Zhang, Cheng, Chao Fan, Wenlin Yao, Xia Hu and Ali Mostafavi. 2019. Social media for intelligent public information and warning in disasters: An interdisciplinary review. *International Journal of Information Management* 49: 190–207.
- Zhu, Xingquan and Xindong Wu. 2004. Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review* 22/3: 177–210.
- Zinsmeister, Heike, Erhard Hinrichs, Sandra Kübler and Andreas Witt. 2009. Linguistically annotated corpora: Quality assurance, reusability and sustainability. In Anke Lüdeling and Merja Kytö eds. *Corpus Linguistics: An International Handbook Vol. 1*. Berlin: Walter de Gruyter, 759–772.
- Zur, Richard M., Yulei Jiang, Lorenzo L. Pesce and Karen Drukker. 2009. Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. *Medical Physics* 36/10: 4810–4818.

Corresponding author

Nicolás José Fernández-Martínez

Department of Languages

Campus de los Jerónimos, Guadalupe 30107

Murcia

e-mail: njfernandez@ucam.edu

received: October 2021

accepted: December 2021

A corpus study of the term *evidence* in open peer reviews to research articles in the *British Medical Journal*

Ingrid García-Ostbye – Barry Pennock-Speck
University of València / Spain

Abstract – The linguistic study of peer-review discourse has focused principally on pre-publication occluded referee reports. However, there are few studies on post-publication open peer reviews of research articles. To address this imbalance, we analyse a type of open peer review, Online Rapid Responses (ORRs) to articles, in the *British Medical Journal* (BMJ), which is the leading medical e-journal. Using a corpus-based approach, we focus on the term *evidence* owing to its importance in scientific discourse. We compiled an *ad-hoc* corpus of 875 ORRs (260,651 tokens) and analysed it using *Wordsmith Tools 6* to ascertain the frequency of *evidence*. We then compared its frequency in our corpus with the *British National Corpus* (BNC), the *Corpus of Contemporary American English* (COCA), the COCA academic subcorpus, the *Cambridge Academic English Corpus* (CAEC) and the sub-corpus of reviews in the *Lancaster-Oslo-Bergen Corpus* (LOB-C). We also performed a keyness analysis of our corpora to ascertain the position of *evidence* and obtained the contexts in which it appears. Our analysis reveals that *evidence* is more frequent in our corpus of ORRs than in general and academic corpora, which highlights its importance in the evaluation of research. Our exploration of its contexts of use show that it reflects the concern of the medical academy for evidence appraisal in state-of-the art medicine.

Keywords – evidence; review; academic writing; computer-mediated communication; rapid response

1. INTRODUCTION

Today, all major scientific journals are peer-reviewed, a practice that dates back nearly 300 years (Paltridge 2017: 22). In the field of medicine, academic peer review originated in the *Philosophical Transactions* of the Royal Society (Räsänen 1999; Mulligan *et al.* 2012). Berkenkotter and Huckin (1999: 62) state that peer review “remains the primary means through which authority and authenticity are conferred upon scientific and scholarly papers by journal editors.” Similarly, Mungra and Webber (2010) also state that peer review bestows authority as well as validity on a published article through a rigorous editorial evaluation process. Despite the importance of peer



review there has been growing criticism of the process due to its tendency to maintain orthodoxy, the lack of specialised reviewers and the poor quality of reviews, as the process constitutes unpaid work and reviewers receive little if any academic recognition (Travis and Collins 1991; Bornmann and Daniel 2008; Thurner and Hanel 2010). Hence, the idea of supplementing the pre-publication (occluded) peer-review process with some form of open post-publication evaluation in the virtual arena is viewed as a way of opening up the vetting process to the entire scientific community in a particular field. Open-access publishing forums are also deemed to increase transparency, address reproducibility issues, improve experimental design, and enhance the analysis of results (Williams *et al.* 2017).

Several medical 2.0 e-journals now include online peer review responses, rapid responses, electronic comments, or e-letters (Hodonu-Wusu 2018), which appear in open post-publication forums or e-journal sections. Unlike occluded peer reviews, the names of the authors of open reviews and their institutions are disclosed to the authors being reviewed and to readers in general. In medicine, these scientific commentaries and e-letters are, nowadays, acknowledged as being crucial in evidence appraisal (Rogers *et al.* 2020).

In particular, Online Rapid Responses (ORRs) to research articles are instances of peer review in the medical academy. They constitute a distinct online medical subgenre in medicine 2.0 and are characterised by their accessibility. Initially, in the *British Medical Journal* (BMJ), they were labelled ‘electronic letters to the editor’ or ‘rapid responses to electronic comments to the editor’. Subscribers can respond to an article by sending a rapid response to the BMJ website.¹ Such responses are freely accessible to readers. This way, peers within the medical community can participate in the online evaluation of a recently published article.

Furthermore, authors of published research articles in the BMJ have the academic duty to respond to any substantive criticism of their papers contained in ORRs. Readers and article authors can opt to receive email updates on the status of their articles, alerting them to corrections and follow-ups by peers and authors, thus emphasising the interactive nature of the website and its evaluation process. ORRs have become established communicative events in the worldwide online medical community. They initiate the post-publication review debate that follows the online publication of a

¹ <https://www.bmj.com/>

research article and complement the occluded review process of the e-journal. This way, the perspective of experts in the field other than journal reviewers or referees is also made available to interested readers.

As a relatively new genre, ORRs have received little attention. To redress this situation somewhat, we have carried out a study of the term *evidence* in a corpus of 875 ORRs (260,651 tokens) extracted from the BMJ, one of the world's foremost medical journals. Using a corpus linguistics approach, we analysed the frequency of *evidence* in our corpus and compared it to other corpora. We also looked at its keyness and the contexts in which it is found. The choice of *evidence* for this study is motivated by the centrality of this term in empirical scientific discourse. Evidence is what separates science from other activities (Husserl 1982; Kuhn 1996). Its importance can be gauged by its presence in the term 'evidence-based medicine', which Sacket *et al.* (1996: 312) define as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients."

As the main aim of this study is to shed light on how open reviewers use the term *evidence* in ORRs and to what purposes, our specific research questions are the following:

- 1) What is the frequency of *evidence* in ORRs? What is its frequency in English varieties such as British English, as represented in the *British National Corpus* (BNC), and American English, as represented in the *Corpus of Contemporary American English* (COCA)? How do they compare?
- 2) What is the frequency of *evidence* in academic English corpora such as the COCA academic subcorpus, the *Cambridge Academic English Corpus* (CAEC) and the subcorpus of reviews of the *Lancaster-Oslo-Bergen Corpus* (LOB-C)?
- 3) Is *evidence* a keyword in the ORRs corpus? What is the position of the term in the keyword list?
- 4) What are the contexts of *evidence* in the corpus of ORRs? What do the colligation and collocation of the term imply regarding the purposes of its usage by open reviewers?

Our research complements previous studies in that it identifies the keywords which characterise ORRs in medicine. Moreover, it also looks in depth into the purpose and usage of the word *evidence*, which is of strategic importance in scientific discourse. Our

study also provides an analysis of the collocational and colligational behaviour of the term in ORRs and adds to existing knowledge regarding this particular online genre.

The paper is organised as follows. In Section 2, we provide a review of the literature on academic reviews and the presence of the term *evidence* in academic corpora. Section 3 deals with the characteristics of our corpus and the methodology used in the study and Section 4 discusses the corpus-based results. Finally, Section 5 provides a summary and some concluding remarks.

2. REVIEW OF LITERATURE

According to the International Committee of Journal Editors (2015), peer review on submitted articles is carried out by experts who are not members of the journal's editorial staff (Hames 2012; Paltridge 2017). Furthermore, peer review is seen as the cornerstone of academic publishing (Hames 2012) and, as such, is essential to the recognition and integration of new research (Hyland 2015).

The discourse of journal reviewers has been the subject of sustained academic study since the 1990s. Several authors have focused on the peer review process and its genres (Kourilová 1996; Okamura and Shaw 2000). Most have centred on 'occluded' reviews, to employ Swales' (1996) term. Occluded peer review and open peer reviews differ at least in two aspects, namely, authorship and readership. The authors of occluded review reports are referees, a select group of experts appointed by the journal editorial board, whereas authors of ORRs are journal readers working in medical institutions, hospitals, or medical research centres worldwide. Regarding readership, occluded peer review reports are only read by authors of research articles, reviewers, and journal editors, whereas ORRs are public and can be read potentially by anyone with access to them.

The study of occluded reviews in different disciplines has mainly focused on their structure, content, and language. Fortanet-Gómez (2008a, 2008b) studied the overall structure of reviewers' reports in the fields of business organisation and applied linguistics. Fortanet-Gómez (2008a: 35) suggests that reviews consist of a four-move structure: (i) summarising judgment regarding suitability for publication, (ii) outlining the article, (iii) points of criticism, and (iv) conclusion and recommendations. Samraj (2016) distinguishes between major revisions and reject reviews, finding patterns

similar to those of Fortanet-Gómez (2008a). She also finds that the two differ organisationally in commentary sections. Paltridge (2017: 41–49), using Fortanet-Gómez’s move structure (2008a), distinguishes between accept reviews, minor revision reviews, major revision reviews, and reject reviews. Paltridge identifies more cases of Move 1 (Judgment Regarding Suitability for Publication) in accept and minor revision reviews and an increasing presence of Move 3 (Points of Criticism) in minor revision reviews, major revision reviews and reject reviews.

Several researchers have focused on the content of occluded referee reports on research articles and have provided the reasons why a paper may or may not be considered worthy of publication. Gosden (2001, 2002, 2003) analyses thematic content in peer reviews and the structure and functions of referees’ comments on scientific papers in chemistry, physics, and microbiology. Of particular interest is that one of the categories identified in referee comments is the consideration of “claims” (Gosden 2003: 92). In this respect, referees often criticise the strength of the researchers’ specific or overall claims, that is, the strength of the evidence they provide. The assessment of the robustness of the researchers’ claims has also been observed in open peer reviews in medicine (García-Ostbye 2018). Woods (2006), acting as a reviewer, categorises the commentaries he makes. His comment categories cover inadequate methods, insufficient explanation of results, limited or misused data, inappropriate choice of journal, problems with presentation and style, unacknowledged bias, inadequate knowledge, limited analysis and inadequate discussion. Coniam (2012), who also acted as a reviewer for the journal *System*, focuses on the areas of the research article that are commented on most frequently: the acceptability of claims, suitability of the methodology, appropriate nature of the data, and clarity of the research questions. Hewings (2004: 260) looks into what was assessed in the reviews to the journal *English for Specific Purposes* (ESP) and concludes that the most commonly evaluated entities are the article itself, expression, claims, analysis, goals, evidence, literature review, bibliographical references, procedure, and knowledge of the field.

Finally, Paltridge (2017: 52–64), also in the field of ESP, looks into positive and negative comments in acceptance and rejection reviews and reviews that ask for minor or major revisions. He tallies the frequency of occurrence of text features in reviews with positive and negative comments, characterising them accordingly. The text features he evaluates are topic, audience, purpose/problem statement/research question, literature

review, methods/research design, presentation and analysis of results, and their discussion/significance, pedagogical implications, language use/style, and clarity. In this respect, a study by Astudillo *et al.* (2016) reveals a higher presence of negative comments in peer review reports.

As for the research article sections that are subjected to most peer review by referees, Belcher (2007) examines positive and negative comments on accepted and rejected papers in reviewers' reports to *ESP*. She identifies the method and discussion sections of research articles as the ones submitted to the most thorough levels of review. Later, Mungra and Weber (2010) study the content of medical review comments, which range from the observation of errors of reasoning regarding the authors' data to the incorrect scientific interpretations of other authors' publications, and the lack of association between data and claims or between claims and prior research. They find that, when assessing a research article in medicine, referees evaluate every section, especially the body of evidence that the author presents in the introduction and discussion sections, the strength of the research method(s) used, their interpretation of results, and the conclusions they reach.

While most studies have focused on the discourse of occluded peer review, fewer have addressed open peer review discourse. One of them is Hyland and Zou (2020: 98), which discusses the extent to which academic blog responses constitute an academic review genre and compares how writers construct criticism in blog responses and book reviews. However, to the best of our knowledge, only García-Ostbye (2018) has addressed ORRs in post-publication medical review fora. She finds more critical responses (61%) than supporting responses (28%) and more infrequent replies to comments (11%). Regarding the structure of ORRs, she characterises it as highly flexible and letter-like. She finds that ORRs include an *ad-hoc* selection of strategies to convince a potentially hostile readership of a personal viewpoint on a published research article, an activity in which the authors' face is central. Finally, she identifies the consideration of evidence in the medical field as one of the strategies in the generic structure of ORRs constituting 24 per cent in the specific medical corpus (24 occurrences), see García-Ostbye (2018: 182).

Regarding the presence of *evidence* in academic corpora, it seems to be affected by corpus sources and discipline-based differences. For example, it is not attested in a relevant position in wordlists from corpora based on research articles and textbooks in

several academic fields: research articles in medicine (Chen and Ge 2007; Wang *et al.* 2008), textbooks in medicine (Hsu 2013), research articles in agriculture (Martínez *et al.* 2009) or in engineering textbooks (Ward 2009; Veenstra and Sato 2018). It is, however, attested among the 100 most frequent words in the research articles in the *Applied Linguistics Research Articles Corpus* (ALC), (see Vongpumivitch *et al.* 2009: 39). Finally, to the best of our knowledge there are no wordlists based on open reviews in medicine except those compiled by García-Ostbye (2018), in which the term *evidence* ranks position 78.

Our research complements previous studies in that it identifies the keywords which characterise ORRs in medicine. Moreover, it also looks in depth into the purpose and usage of the word *evidence*, which is of strategic importance in scientific discourse. Our study also provides an analysis of the collocational and colligational behaviour of the term in ORRs and adds to existing knowledge regarding this particular online genre.

3. CORPUS CHARACTERISTICS AND METHODOLOGY

The main criterion for selecting the BMJ as our corpus is its worldwide prestige. It is one of the leading online journals in the medical field, with an impact factor of 30.313 for the period between 2019 and 2020.

The texts that make up the corpus represent authentic online discourse in medicine, authored by medical professionals and researchers in medicine worldwide. As such, these texts are examples of expert-to-expert communication in English as a lingua franca. The corpus comprises all 875 ORRs to research articles published on the BMJ website² in 2006. These contained 260,651 tokens (word forms, digits, abbreviations regardless of how often they are repeated) and 13,873 types (distinct word forms, digits or abbreviations). The year 2006 was chosen as access was free to non-subscribers.

To ascertain the weight and importance of the term *evidence* in our corpus, we first compared its normalised frequency with that found in the academic component of the BNC and COCA. We also compared the normalised frequency of *evidence* in our corpus with that attested in the *Cambridge Academic English Corpus* (CAEC), which was accessed through *SketchEngine* (Kilgariff *et al.* 2014),³ and with a sub-corpus of

² <https://www.thebmj.com/>

³ <https://www.sketchengine.eu/bibliography-of-sketch-engine/>

reviews from the *Lancaster-Oslo-Bergen* Corpus (LOB-C). Then, we ran a keyword search, a Keyword in Context (KWIC) search and a collocation and colligation search using *Wordsmith Tools 6* (Scott 2012)

4. RESULTS AND DISCUSSION

4.1. Frequency of the term *evidence* in ORRs and other corpora

We will first present our results for the normalised frequency of *evidence* in our corpus compared to the others (see Table 1 below).

Corpus	Normalised frequencies
ORRs	1.370
BNC	0.215
COCA	0.161
LOB-C	0.001
CEAC	0.498
COCA (academic subcorpus)	0.236

Table 1: Normalised frequencies per 1,000 words of *evidence* in six corpora

As can be seen in Table 1, the low frequency for *evidence* in the general English corpora was unsurprising as this term would only be expected to occur in a limited number of contexts in general English. We also thought it would be unlikely that *evidence* would be frequent in the LOB-C specialised review corpus as it was made up of non-academic texts. We did not expect, however, that the difference between the relativised frequency of *evidence* in our corpus compared to the academic subcorpora in CEAC and COCA would be so great. Interestingly, after the removal of common function words, using a stoplist, the term has been observed among the most common words occurring at least 1,000 times in occluded peer review reports in the *BMJ* (Falk Delgado *et al.* 2019).

4.2. ORR corpus keywords

In the next stage of our analysis, we ran the keyword search option in *Wordsmith Tools 6* using the LOB-C corpus as our reference corpus. The term *evidence* was found to be the 21st most frequent keyword, as shown in Table 2 below. Regarding the function words in the table, they show some of the characteristics of scientific discourse. For example, the word *that*, which can represent a conjunction, a demonstrative an adjective, a demonstrative pronoun or a relative pronoun, is the most frequent keyword. The third most frequent keyword, *we*, indicates the presence of research authors and

appeals to the scientific community (namely, health professionals in academia) in these post-publication online debates and highlights their social nature. The first-person plural pronoun helps establish interpersonal relationships with the reader (Hyland 2005) and thus emphasises medical activity as a communal endeavour. The high frequency of *we* suggests an egalitarian online relationship and such a result contrasts sharply with previous studies, such as that of Samraj (2021), which singled out second person pronouns as a discourse characteristic of negative evaluation in occluded manuscript reviews. Regarding the high frequency of *may* and *would*, it may be explained by the presence of tentative explanations, discussions, and hedged alternative interpretation of research results, which are characteristic of scientific discourse in medicine (Salager-Meyer 1994).

	Keyword	Frequency in the ORRs Corpus	Percentage in the ORRs Corpus	Number of texts where it occurs
1	<i>That</i>	3,567	1.37	303
2	<i>Study</i>	1,242	0.48	11
3	<i>We</i>	1,202	0.46	43
4	<i>Patients</i>	1,193	0.46	0
5	<i>Would</i>	830	0.32	39
6	<i>May</i>	824	0.32	22
7	<i>Risk</i>	719	0.28	1
8	<i>Et</i>	558	0.21	0
9	<i>Al</i>	545	0.21	0
10	<i>Results</i>	493	0.19	1
11	<i>Health</i>	489	0.19	1
12	<i>Treatment</i>	488	0.19	7
13	<i>Trial</i>	476	0.18	1
14	<i>Authors</i>	476	0.18	4
15	<i>Data</i>	419	0.16	0
16	<i>Care</i>	418	0.16	2
17	<i>Group</i>	416	0.16	8
18	<i>Studies</i>	402	0.15	5
19	<i>Analysis</i>	367	0.14	2
20	<i>Research</i>	363	0.14	3
21	<i>Evidence</i>	351	0.14	5

Table 2: Keywords in ORRs

As we expected, the most frequent content words are all related to the fields of medicine and research: five words for medicine and 12, including *evidence*, for research. In the area of medicine, we have *patients*, *risk*, *health*, *treatment*, and *care*. The words related to research comprise *study*, *et*, *al*, *results*, *authors*, *trial*, *data*, *group*, *studies*, *analysis*, *research*, and *evidence*.

4.3. Contexts of the keyword evidence

To ascertain how the term *evidence* is used and why, we analysed the most frequent contexts of the term in the corpus of ORRs. The contexts were identified and categorised inductively, focusing on their collocation and colligation. In what follows, we provide examples of *evidence* in these contexts.

4.3.1. Negative particles preceding evidence

The term *evidence* is often preceded by negative particles in ORRs (38 occurrences), generally the negative determiner *no*, to signal insufficient or complete lack of evidence to make an assertion in the field. Table 3 below includes some illustrative examples.

EXAMPLES	
1.	In response to the recommendation of the delayed prescribing approach for acute infective conjunctivitis, it is indeed a novel idea for reducing medicalisation of many self-limiting disease in the community. There is no current evidence that routine use of topical antibiotics reduces the bacterial load in the community.
2.	Compared with immediate antibiotics delayed prescribing had the advantage of reduced antibiotic use, no evidence of medicalisation , similar symptom control, and reduced reattendance for eye infections.
3.	In conclusion, residents of major metropolitan areas live in a veritable sea of radio-frequency energy. Despite this, there is no epidemiological evidence that continuous exposure to low amounts of electromagnetic energy plays any role in causing cancer.
4.	Radio operators have been sensitized to be alert to the possibility of medical consequences caused by unsafe transmitter operation. Yet with all of these warnings, there is no evidence – anecdotal or otherwise – that a physical presence near low power radio transmitters or antennae causes any adverse medical consequence to Amateur Radio operators.
5.	The Buscemi and colleagues' meta-analysis (1) concludes that there is no evidence that melatonin is effective in treating secondary sleep disorders or sleep disorders accompanying sleep restriction. Here, readers need to recall that there is another meta-analysis by these colleagues, pointing at the efficacy and safety of melatonin in the management of chronic or primary insomnia (2)
6.	With a RR of 0.65 and 95% CI of 0.48 to 0.88 from like-with-like cohort studies, I submit that to put out the message that there is no clear evidence to support a reduction in risk of CV death from long chain omega-3 usage is highly irresponsible.
7.	Compared to patients operated on within 24 hours, delay to surgery in patients who were initially medically unfit was associated with increased mortality (hazard ratio 1.3; 95% confidence interval 1.1 to 1.4). However, there was no evidence of an association between delay to surgery and mortality for patients whose operation was delayed for administrative reasons (HR 0.9, 95% CI 0.8 to 1.0) or for other reasons (HR 1.1, 95% CI 0.9 to 1.2).

Table 3: Some examples of negative determiners preceding *evidence* in ORRs

As shown in Table 3, the negative determiners preceding *evidence* highlight the presence of criticism in ORRs. This is to be expected as negative appraisal is an important function of any type of review (Samraj 2016). In this respect, they resemble occluded reviews of manuscripts that are deemed to require major revision or are

rejected, which is consistent with previous research on occluded peer review referee reports (Kourilovà 1996; Samraj 2016). Criticism has also been observed by Hyland and Zou (2020) in academic blog responses and by Rogers *et al.* (2020) in critical appraisal in letters.

4.3.2. Colligation of *evidence* with premodifiers

In ORRs, the term *evidence* colligates with adjectives (108 occurrences) which refer to the presence, amount, quality, and nature of the evidence provided in the article under open review and/or the evidence available in the field of medicine, as illustrated in Table 4 below.

EXAMPLES	
1.	There was also insufficient evidence to evaluate the accuracy of MRI in patients presenting with different clinical symptoms.
2.	MR was not withdrawn on the basis of Wakefield's anecdotal evidence in 12 patients -instead it was subjected to proper scrutiny.
3.	If authors refer specifically to older patients, clear evidence is derived from the MEDENOX (3) and PREVENT (4) studies.
4.	Harnden <i>et al.</i> present convincing evidence that a common cause (Bordetella pertussis) of persistent cough in adolescents and adults also extends to school age children
5.	There is no doubt about the emerging evidence demonstrating that acupuncture may have some specific treatment efficacy in knee pain1, neck pain 2 and back pain 3.
6.	There is little doubt that increasing evidence is emerging which indicates that the context in which a treatment is delivered may be of great importance 9. If this is indeed the case, then the major effects which we observe within clinical trials of both acupuncture and conventional medicine may be more related to the context and environment of the trial than with the specific efficacy of the treatment being studied 6, 9.
7.	For alpha-linolenic acid, the epidemiological evidence is less convincing and randomized controlled trials are lacking.
8.	The reason why indication for reconstruction is not clear cut is that we lack scientific evidence that it prevents from late osteoarthritis, and this is the second point we would like to discuss.

Table 4: Some examples of colligation of the term *evidence* with premodifiers in ORRs

Adjectives that denote low quantities justify the relativisation or rejection of a particular claim or assertion in the article or the discipline, such as, for example, *insufficient* (2 occurrences), *limited* (2 occurrences), *poor* (1 occurrence), *anecdotal* (1 occurrence). On the contrary, there are adjectives that express the writer's support of the evidence existing in the field or provided in the paper for example, *clear* (6 occurrences), *strong* (4 occurrences), *convincing* (3 occurrences), *valid* (2 occurrences), *good* (1 occurrence), *abundant* (1 occurrence), *best available* (1 occurrence), *current* (1 occurrence), *important* (1 occurrence). Some adjectives, such as *emerging* (3 occurrences), *growing* (2 occurrences), or *increasing* (1 occurrence), refer to changes in the amount of

evidence analysed and reveal emerging research tendencies in the discipline. Other adjectives, such as *epidemiological* (3 occurrences), *experimental* (1 occurrence), *scientific* (5 occurrences), and *randomised* (1 occurrence), reveal the origin, type, or category of evidence, which helps the virtual community focus on these particular research categorisations.

4.3.3. Collocation of *evidence* with postmodifying non-finite noun clauses

Evidence is often followed (26 occurrences) by postmodifying non-finite noun clauses (Biber *et al.* 1999: 291). In Table 5 below, the first three examples refer to the person or research group providing the evidence, the next three allude to the availability of the evidence, and the last two indicate the conclusions to be drawn from the evidence.

EXAMPLES	
1.	The evidence cited by Lewith and White is unpersuasive or totally not applicable. ... If there is no placebo control possible for acupuncture experiments then it becomes impossible to ever falsify acupuncture claims and the proposition that acupuncture is more than placebo
2.	However, the evidence presented by Bekkelund and colleagues for visual field loss is, for several reasons, far from convincing . The authors do not represent the visual field in the most appropriate format for interpretation and they do not provide any indication as to the reliability of the response from the patient during each examination.
3.	The evidence provided by the paper however appears incomplete . The definition of quality is performance in the clinical domains of the new GMS contract. However, this only measures part of the quality spectrum for primary care activity.
4.	In the introduction to their paper, ..., the authors also state that the baseline risk of VTE in medical patients remains uncertain as does the effectiveness of thromboprophylactic therapy in these individuals. Neither of these statements can be substantiated on review of the evidence available regarding the risk of VTE in acute medical admissions and the established thromboprophylactic therapies used in these patients
5.	The assertion of Dr Antony in his rapid reply that “some writers and researchers have concluded that these dangers are so great that online discussion groups should be professionally moderated.” Is not referenced and I doubt that anybody actually said something like this. I am not aware of any evidence suggesting that moderated communities are “better” than unmoderated communities, or the other way around [2]. The Esquivel paper unfortunately doesn’t contribute to answering this question.
6.	Would there be unexpected repercussions, including retinopathy, when hyperbaria is combined with 100% oxygen that too when the blood-brain barrier is compromised in HIE? May I remind that evidence supporting resuscitation of newborn with room air (not 100% oxygen) is mounting in the recent western 4 and eastern 5 literature .
7.	Surely with abortions in the UK now exceeding 200 000 a year, the evidence indicating a link with preterm births deserved at least a mention or else evidence that there is no such association should have been presented to reassure the thousands of women undergoing terminations of pregnancy each week?
8.	We must try to suspend our own opinion until there is evidence regarding efficacy of Japanese acupuncture which is widely legitimized. There are therefore a number of other possible explanations for the study. The first is that all acupuncture may be placebo with deeper and more painful needling being a more powerful placebo than superficial needling

Table 5: Collocation of the term *evidence* with postmodifying non-finite noun clauses

After a close reading of the examples from the corpus, we have identified that an important function of these postmodifying clauses is to express criticism (e.g., *evidence cited by [...] is unpersuasive*, *evidence presented by [...] is [...] far from convincing*, *evidenced provided by [...] appears incomplete*). These critical acts may constitute a direct attack on the researchers' reasons for making assertions, for legitimising their claims, and may also be considered face-threatening acts. These clauses are also found more often simply to provide a description of the evidence (e.g., *evidence available*, *evidence suggesting*, *evidence supporting*, *evidence indicating*, *evidence regarding*), as can be seen in examples 4 to 7 in Table 5.

4.3.4. Colligation of *evidence* with *that*-clauses

As we can see in Table 6, *evidence* colligates with *that* complement clauses (47 occurrences). As was the case with postmodifying non-finite noun clauses, their main function is to provide more information about the *evidence* being mentioned such as, for instance, in example 1 (see Table 6) where the effect of low molecular weight heparins on bleeding complications is mentioned. Similarly, in example 2, information is added regarding proof that a common cause of persistent cough is found not only in adolescents and adults but also in school children, or, in example 3, where information on the relation between osteonecrosis and low-dose corticosteroids is provided.

EXAMPLES	
1.	The results of Cook <i>et al.</i> conflict with the evidence that the use of low molecular weight heparins (LMWH) for thromboprophylaxis does not cause more bleeding complications than heparin (UFH).
2.	Harnden <i>et al.</i> present convincing evidence that a common cause (Bordetella pertussis) of persistent cough in adolescents and adults also extends to school age children.
3.	However, we do not agree that the references you quoted provide evidence that osteonecrosis was related to low-dose corticosteroids.
4.	The Buscemi and colleagues meta-analysis (1) concludes that there is no evidence that melatonin is effective in treating secondary sleep disorders or sleep disorders accompanying sleep restriction. Here, readers need to recall that there is another meta-analysis by these colleagues, pointing at the efficacy and safety of melatonin in the management of chronic or primary insomnia.
5.	First, without any scientific evidence McCrea (2) and Nesbitt (3) responses try to convince the readers that hip fracture precedes the fall (breaking and falling) and not vice versa (falling and breaking). However, the landmark study on injury mechanisms of hip fracture (4), followed by many others (5,6,7), has given strong evidence that majority of hip fractures among older adults are caused by a sideways fall onto the hip (greater trochanter).
6.	Finally, the rate of fractures outside hip and pelvis was similar in our protector group and control group thus providing further evidence that the groups were very comparable.

Table 6: Colligation of *evidence* with *that*-clauses

4.3.5. Colligation of *evidence* with premodifiers and *to*-infinitive clauses

In our corpus, *evidence* is also attested in a syntactic relation with premodifiers and *to*-infinitive clauses (29 occurrences). The number of examples displaying a negative appraisal of *evidence* with a *to*-infinitive clause is 15, while there are 12 examples showing positive appraisal. Whether appraisal is positive or negative depends on the premodifiers. As can be seen in Table 7, the first four examples collocate with *evidence* to show its strength while the remaining examples show its weakness.

EXAMPLES	
1.	CT head exposes a child to a dose of 0.5Gy (approx. 100 chest X-rays) (3) and there is some evidence to show a detrimental effect on children's cognitive development (4).
2.	There is a body of evidence to suggest that Sudden Infant Death Syndrome commonly referred to as cot death is linked to a mild diffuse brain injury acquired by the fetus in utero or during the birth procedure linked to trauma.
3.	We do not think that there is sufficient evidence from RCTs to convince us of the safety in a re world population undergoing hysterectomy.
4.	Hence a longitudinal study to follow up on the present result to observe the association between these two factors with other social and economic environment would give health care providers comparatively strong evidence to predict to a certain extent one's health outcome given a person is both a smoker and is obese.
5.	The fact that previous Cochrane reviews by Hooper have concluded that there was no evidence to support the hypothesis that reducing dietary salt intake decreased blood pressure, or that reducing saturated fat intake decreased heart disease risks seems to me to make the point very eloquently.
6.	Thus, there is no valid evidence to recommend perioperative beta-blockade on the sole indication of diabetes mellitus.
7.	The quality of health care has improved immeasurably in recent years because of the combined efforts of doctors and researchers to use the best available evidence to inform decisions about clinical practice for individual patients .
8.	There was also insufficient evidence to evaluate the accuracy of MRI in patients presenting with different clinical symptoms.

Table 7: Colligation of *evidence* with premodifiers and *to*-infinitive clauses

The most common verbs in the *to*-infinitive clauses colligating with *evidence* are *support* (5 occurrences), *suggest* (9 occurrences), *understand* (2 occurrences), and *show* (2 occurrences). The remaining verbs are only found once (e.g., *change*, *evaluate*, *extract*, *guide*, *inform*, *justify*, etc.). Infinitives are employed to indicate some of the uses the medical community makes of the evidence at their disposal (e.g., *to show effects*, *to suggest*, *to convince*, *to support a hypothesis*, *to make predictions or recommendations*).

4.3.6. Collocations of *evidence* with partitive noun phrases

On the basis of the data from the open review corpus, *evidence* also collocates with partitive noun phrases (22 occurrences). Partitive noun phrases assess the presence or absence of evidence in the field, its quantification, qualification, classification, or the indication of its sources. They also signal support for the assertions made about the evidence, as exemplified in Table 8.

EXAMPLES	
1.	There is a body of evidence to suggest that Sudden Infant Death Syndrome commonly referred to as cot death is linked to a mild diffuse brain injury acquired by the fetus in utero or during the birth procedure linked to trauma.
2.	Indeed, there is a wealth of evidence from published studies to show that supported quit attempts are much more successful than unsupported ones. (Tobacco smoking cessation)
3.	There is quite a bit of evidence from Table 3 that the two groups were not similar at baseline and the lack of a major impact may, as in the early Head Start evaluations, be due to confounding.
4.	This highlights the importance of using the totality of evidence from other trials.
5.	The bias came from evaluating only randomized clinical trials (RCTs), omitting large amounts of published evidence (2) on how dietary omega-3 fats compete with omega-6 fats as they maintain healthy tissues and prevent disease processes.
6.	Few other interventions-lyfestyle, pharmacologic, or surgical have levels of evidence or magnitudes of health benefits approaching those of fish consumption. For example, in a meta-analysis of 14 randomized trials of statin therapy, considered by many a pharmacologic panacea, total mortality was reduced by 12 % (RR=0.88,95% CI=0.84-0.91).
7.	Meta-analysis is considered the highest order of evidence and the definitive source of conclusive data. Ho and Sheridan [1] failed to report on a more practical and common side effect of prolonged intravenous frusemide use.
8.	We would concur with the concerns expressed by Neely <i>et al.</i> (5) about the lack of evidence of the superiority of tinzaparin over UFH.

Table 8: Collocations of *evidence* with partitive noun phrases

Partitive noun phrases reveal the presence of an ongoing process of evidence appraisal, with sometimes conflicting evidence, which may directly impact decision processes, risk taking, and cost/benefit assessments adopted by medical professionals in patient care and hospital settings, among others.

4.3.7. Colligation of *evidence* with preceding verbs

The data retrieved from the ORRs corpus shows that *evidence* colligates in the same clause with preceding verb phrases (71 occurrences). There is a great variety of preceding verbs but the most common ones refer to the evidence provided or used; these include *provide* (15 occurrences), *show* (3 occurrences), *give* (3 occurrences), *offer* (2 occurrences) and *use* (3 occurrences). Table 9 below includes some illustrative examples. As shown, two of them, namely *ignore* and *disregard*, refer to obviating the

evidence (see examples 6 and 7), while example (8), *rests on*, refers to the lack of credibility of the evidence presented by researchers.

EXAMPLES	
1.	In any case they do not offer any evidence to support this statement.
2.	Everitt <i>et al.</i> have presented an interesting article showing evidence that delayed antibiotic prescribing is the best management strategy in acute infective conjunctivitis (AIC).
3.	The paper by Zackrisson <i>et al.</i> (1), reporting the follow-up of the Malmo trial, provides important evidence .
4.	However, the landmark study on injury mechanisms of hip fracture (4), followed by many others given strong evidence that majority of hip fractures among older adults are caused by a sideways hip (greater trochanter).
5.	Several large double-blinded placebos controlled antibiotic trials have attempted to clarify the role of C pneumoniae in CHD with conflicting results (3). Only a few trials have used evidence of pneumoniae infection as inclusion criteria (and these were dependent on antibody measurements).
6.	The context presented by Hooper <i>et al.</i> ignores strong biological evidence for the potentially disease-specific effects of omega-3 fat.
7.	As it would not be right to accept the results of the previous meta-analysis as the end of the argument, it would be equally incorrect to disregard all the evidence supporting metformin plus clomiphene treatment published before the findings of Moll <i>et al.</i> based on the trial alone.
8.	Hooper <i>et al.</i> 's condensed version (1) of their 2004 Cochrane omega 3 review up to February 2002 (2) concluded that it is not clear that omega 3 fats alter total mortality... We do not believe that this conclusion rests on solid evidence (1).

Table 9: Colligation of *evidence* with preceding verbs

Although many of these verbs serve the purpose of praising or criticising the research articles under scrutiny, their presence also reveals how agile these online debates between researchers are (e.g., *present*, *summarise*, *cite*, *used*), how medical community members operate around the concept of *evidence* (e.g., *showing*, *disregard*, *ignores*), and how its treatment is flexible (e.g., from conclusions resting on solid evidence and presenting convincing evidence to arbitrarily ignoring or disregarding evidence).

4.3.8. Colligation of *evidence* with following verbs

Most of the verbs following *evidence* (28 occurrences) refer to what it reveals overtly or tentatively: *suggest* (6 occurrences), *support* (4 occurrences), *present* (2 occurrences), *provide* (2 occurrences), *show* (2 occurrences), *indicates* (1 occurrence), *points to* (1 occurrence), *produce* (1 occurrence). Some of these verbs are shown in Table 10 below. Two of them, *have* (3 occurrence) and *obtain* (1 occurrence), refer to the evidence the researchers possess (examples 9 and 10). The remaining verbs have diverse meanings: *appear* (1 occurrence), *arise* (1 occurrence), *change* (1 occurrences), *emerge* (1 occurrence), and *fulfil* (1 occurrence).

EXAMPLES

1. **The clinical evidence would suggest** there is a difference in effect size between Streitberger and real acupuncture 7, but it may just be that Kaptchuk's placebo is a less effective form of acupuncture that is practised in Japan.
 2. Radiation must certainly be taken into account when balancing risk versus benefit in use of CT in mild head injury. **Current best evidence suggests** that the risk of an occasional CT is low, but special caution is needed in children below 18 months.
 3. Our report on the DIPOM trial (1) does not state that metoprolol is of no benefit peri-operatively in non-cardiac surgery. As suggested by McCulloch, we humbly state that **no evidence supports this intervention**.
 4. Jonathan T McCrea makes an interesting point about how hip fractures happen, although as far as I know **the evidence points to** the impact of a fall as the usual cause.
 5. Also, **the evidence could change** rapidly when data from further RCTs at low risk of bias become available.
 6. **Evidence would arise** from a statement like this: "In a consecutive review of all cases delivered to our unit, we found 15 patients with osteonecrosis who received and A cases who did not receive low-dose corticosteroids."
 7. If new **evidence fulfils** our inclusion criteria, we will be happy to include it when updating our review in the future.
 8. I do not claim that osteonecrosis is not associated with low-dose corticosteroids. I only say that **evidence should be obtained** by appropriate methods.
-
-

Table 10: Colligation of *evidence* with following verbs

There is a great abundance of hedged statements among the verbs following *evidence*. The verb *suggest* itself, which is found six times, is intrinsically non-categorical. Hedging is also expressed through conditionals (e.g., *would suggest*, *if new evidence fulfils*, *suggests*, *would arise*, *could change* and phrases such as *as far as I know*). This lack of assertiveness seems to support the adage that no research report provides proof.

5. CONCLUDING REMARKS

This study is the first to focus on ORRs in any electronic journal and provides an insight into the nature of this type of open review, thus complementing studies on traditional occluded peer-reviews. Although it is only the 21st most frequent word in our corpus of ORRs, we have focused on the term *evidence* as, more than any other, it epitomises empirical research, which has come under attack from certain sectors of society (Nasr 2021). Detractors of scientists and science often criticise the opaqueness of empirical research, mainly carried out in laboratories and field studies that include double-blind trials and other procedures which follow strict protocols of confidentiality. These critics also decry what they see as the unquestionable nature of evidence. ORRs provide transparency regarding research methods and results and highlight the contested nature of what is considered *evidence* in medicine. In this sense, ORRs reflect their authors' concern that the research they are assessing should provide sufficient evidence to

convince the reader and the academy and that it builds on present existing research in the field in even more detail.

In our study, we have ascertained that the term *evidence* is not only used more often in ORRs than in other more general corpora, but it is also a keyword of strategic importance in the medical field. The study of how the term is used in ORRs leads us to conclude that the evidence provided by the researchers as well as their claims and assertions regarding the state of the art in the discipline are carefully scrutinised not only by editorial teams and referees, but also by online medical peer experts in the virtual arena before they can become accepted knowledge in the discipline.

Through open peer review, this process becomes even more thorough as it addresses the reliability and validity of the published paper and the most up-to-date research in the field in even more detail, and this has become a characteristic feature of medicine 2.0 culture.

Our examination of words and phrases that collocate with *evidence* shows that, on the one hand, it is preceded by negative particles, determiners and adjectives, partitive noun phrases and verbs and, on the other, followed by postmodifying non-finite noun clauses, *that*-clauses, and *to*-infinitive clauses. By examining the collocations of the term *evidence* in ORRs, we have pinpointed the various ways reviewers assess whether article authors have fully explored the latest developments in the field. We can also see how reviewers examine the inferential processes implemented by the authors when making claims and assertions. The critical nature of reviews becomes apparent through the large number of negative particles that collocate with *evidence*. What is also evident is the frequently tentative nature of the reviewers' comments seen in the presence of hedging devices.

The observation of the term *evidence* in its various contexts places it strategically in the centre of the evaluation process performed by medical experts. In the experts' appraisal of research articles, medical evidence is not considered absolute proof; instead, it points in a particular direction. The careful choice of the words surrounding *evidence* helps to establish the degree of probability of truthfulness of assertions or propositions in the discipline and to what extent the online community of medical professionals can believe them and rely on them for medical practice. That is the reason why medical peers carefully scrutinise the incorporation of new knowledge into the discipline and ensure it reaches the highest scientific standards.

This study has revealed that, in ORRs, the term *evidence* refers to cutting-edge medical knowledge widely recognised by the medical community. It also refers to relevant research in the medical field whose conclusions and assertions have been accepted by the medical community because they have been obtained by adequate experimental design, appropriate methods, and logical, scientific reasoning. In the corpus of ORRs, the term *evidence* refers to both the research article under open review and to the state of the art in the medical field.

This study suggests the presence of an ongoing process of evidence appraisal in academia, in the virtual arena, where experts praise or criticise articles on the grounds of the evidence needed to make assertions and emerging research tendencies. The wide range of uses of this term in the online debates in the BMJ implies that evidence is at the heart of medical academy activity.

There are two main limitations to our study, which, in turn, suggest future avenues of research. The first is that we centre on one medical e-journal, the BMJ. Further studies are needed to discern whether the analysis of *evidence* in different medical e-journals would yield similar results and whether differences would arise across disciplines. Second, as we focus exclusively on the term *evidence*, it would be enlightening to include content words, such as *study*, *patients*, *risk*, or *results* found in our keyword list.

REFERENCES

- Astudillo, César, Karem Squadrito, Germán Varas, Carlos González and Omar Sabaj. 2016. Polarity of comments and internal consistency in peer review reports on scientific research articles. *Acta Bioethica* 22/1: 119–125.
- Belcher, Diane D. 2007. Seeking acceptance in an English-only research world. *Journal of Second Language Writing* 16/1: 1–22.
- Berkenkotter, Carol and Thomas N. Huckin. 1999. *Genre Knowledge in Disciplinary Communities*. Mahwah: Lawrence Erlbaum Associates.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.
- Bornmann, Lutz and Hans-Dieter Daniel. 2008. Selecting manuscripts for a high-impact journal through peer review: A citation analysis of communications that were accepted by *Angewandte Chemie International Edition* or rejected but published elsewhere. *Journal of the American Society for Information Science and Technology* 59/11: 1841–1852.
- Chen, Qi and Guang C. Ge. 2007. A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles. *English for Specific Purposes* 26/4: 502–514.

- Coniam, David. 2012. Exploring reviewer reactions to manuscripts submitted to academic journals. *System* 40/1: 1–28.
- Falk Delgado, Alberto, Gregori Garretson and Anna Falk Delgado. 2019. The language of peer review reports on articles published in the BMJ, 2014–2017: An observational study. *Scientometrics* 120/3: 1225–1235.
- Fortanet-Gómez, Inmaculada. 2008a. Evaluative language in peer review referee reports. *Journal of English for Academic Purposes* 7/1: 27–37.
- Fortanet-Gómez, Inmaculada. 2008b. Strategies for teaching and learning an occluded genre: The RA referee report. In Sally Burgess and Pedro Martín-Martín eds. *English as an Additional Language in Research Publication and Communication*. Bern: Peter Lang, 19–38.
- Garcia-Ostbye, Ingrid K. 2018. *Electronic Responses from the Online British Medical Journal: A Case Study in Genre Analysis*. Beau Bassin: Editorial Académica Española.
- Gosden, Hugh. 2001. ‘Thank you for your critical comments and helpful suggestions’: Compliance and conflict in authors’ replies to referees’ comments in peer reviews of scientific research papers. *Iberica* 3: 3–17.
- Gosden, Hugh. 2002. Thematic content in peer reviews of scientific papers. *Seminarios de Lingüística* 5: 56–75.
- Gosden, Hugh. 2003. ‘Why not give us the full story?’ Functions of referees’ comments in peer reviews of scientific research papers. *English for Specific Purposes* 22/1: 87–101.
- Hames, Irene. 2012. Peer review in a rapidly changing landscape. In Robert Campbell, Ed Pentz and Ian Borthwick eds. *Academic and Professional Publishing*. Cambridge: Chandos Publishing, 15–52.
- Hewings, Martin. 2004. An ‘important contribution’ or ‘tiresome reading’? A study of evaluation in peer reviews of journal article submissions. *Journal of Applied Linguistics* 1/3: 247–274.
- Hodonu-Wusu, James O. 2018. Open science: A review on open peer review literature. *Library Philosophy and Practice*: 1–19. <https://digitalcommons.unl.edu/libphilprac/1874/>
- Hsu, Wenhua. 2013. Bridging the vocabulary gap for EFL medical undergraduates: The establishment of a medical word list. *Language Teaching Research* 17/4: 454–484.
- Husserl, Edmund. 1982. *Ideas Pertaining to a Pure Phenomenology and a Phenomenological Philosophy. First Book: General Introduction to a Pure Phenomenology*. The Hague: Nijoff.
- Hyland, Ken. 2005. Representing readers in writing: Student and expert practices. *Linguistics and Education* 16/4: 363–377.
- Hyland, Ken. 2015. *Academic Publishing: Issues in the Challenges in the Construction of Knowledge*. Oxford: Oxford University Press.
- Hyland, Ken and Hang Zou. 2020. Managing evaluation: Criticism in two academic review genres. *English for Specific Purposes* 60: 98–112.
- International Committee of Medical Journal Editors. 2015. *Responsibilities in the Submission and Peer-review Process*. <http://www.icmje.org/recommendations/browse/roles-and-responsibilities/responsibilities-in-the-submission-and-peer-peview-process.html> (9 July, 2021.)

- 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.
- Kourilová, Magda. 1996. Interactive functions of language in peer reviews of medical papers written by non-native users of English. *UNESCO ALSED-LSP Newsletter* 19/1: 4–21.
- Kuhn, Thomas S. 1996. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Martínez, Iliana A., Silvia C. Beck and Carolina B. Panza. 2009. Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes* 28/3: 183–198.
- Mulligan, Adrian, Louise Hall and Ellen Raphael. 2012. Peer review in a changing world: An international study measuring the attitudes of researchers. *Journal of the American Society for Information Science and Technology* 64/1: 132–161.
- Mungra, Philippa and Pauline Webber. 2010. Peer review process in medical research publications: Language and content comments. *English for Specific Purposes* 29/1: 43–53.
- Nasr, Nancy. 2021. Overcoming the discourse of science mistrust: How science education can be used to develop competent consumers and communicators of science information. *Cultural Studies of Science Education* 16/2: 345–356.
- Okamura, Akiko and Phillip Shaw. 2000. Lexical phrases, culture and subculture in transactional letter writing. *English for Specific Purposes* 19/1: 1–15.
- Paltridge, Brian. 2017. *The Discourse of Peer Review: Reviewing Submissions to Academic Journals*. London: Palgrave Macmillan.
- Räsänen, Christine. 1999. *The Conference Forum as a System of Genres*. Göteborg: Acta Universitatis Gothoburgensis.
- Rogers, James R., Hollis Mills, Lisa V. Grossman, Andrew Goldstein and Chunhua Weng. 2020. Understanding the nature and scope of clinical research commentaries in PubMed. *Journal of the American Medical Informatics Association* 27/3: 449–456.
- Sackett, David L., William M. C. Rosenberg, John A. Muir Gray, R. Brian Haynes and W. Scott Richardson. 1996. Evidence based medicine: What it is and what it isn't. *BMJ* 312: 71–72.
- Salagar-Meyer, Françoise. 1994. Hedges and textual communicative function in medical English written discourse. *English for Specific Purposes* 13: 149–170.
- Samraj, Betty. 2016. Discourse structure and variation in manuscript reviews: Implications for genre categorization. *English for Specific Purposes* 42: 76–88.
- Samraj, Betty. 2021. Variation in interpersonal relations in manuscript reviews with different recommendations. *English for Specific Purposes* 62: 70–83.
- Scott, Michael. 2012. *WordSmith Tools 6*. Stroud: Lexical Analysis Software.
- Swales, John M. 1996. Occluded genres in the academy: The case of the submission letter. In Eija Ventola and Anna Mauranen eds. *Academic Writing: Intercultural and Textual Issues*. Amsterdam: John Benjamins, 45–58.
- Turner, Stefan and Rudolf Hanel. 2010. Peer-review in a world with rational scientists: Toward selection of the average. *European Physical Journal* 84/4: 707–711.
- Travis, G. David and Harry M. Collins. 1991. New light on old boys: Cognitive and institutional particularism in the peer review system. *Science, Technology, & Human Values* 16/3: 322–341.

- Veenstra, Jay and Yoko Sato. 2018. Creating an institution-specific science and engineering academic word list for university students. *Journal of Asia* 15/1: 148–166.
- Vongpumivitch, Viphavee, Ju-yu Huang and Yu-Chia Chang. 2009. Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes* 28/1: 33–41.
- Wang, Jing, Liang Shao-lan and Guan-Chun Ge. 2008. Establishment of a medical academic word list. *English for Specific Purposes* 27/4: 442–458.
- Ward, Jeremy. 2009. A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes* 28/3: 170–182.
- Williams, Michael, Kevin Mullane and Michael J. Curtis. 2017. Addressing reproducibility: Peer review, impact factors, checklists, guidelines, and reproducibility initiatives. In Michael Williams, Kevin Mullane and Michael J. Curtis eds. *Research in the Biomedical Sciences, Transparent and Reproducible*. San Francisco: Elsevier Academic Press, 197–306.
- Woods, Peter. 2006. *Successful Writing for Qualitative Researchers*. London: Longman.

Corresponding author

Ingrid García-Ostbye
 Facultat de Filologia, Traducció i Comunicació
 Departament de Filologia Anglesa i Alemanya
 Avinguda Blasco Ibañez 32
 Valencia 46010
 Spain
 E-mail: ingrid.k.garcia@uv.es

received: September 2021
 accepted: February 2022

Preterit-imperfect acquisition in L2 Spanish writing: Moving beyond lexical aspect

Sophia Minnillo – Claudia Sánchez-Gutiérrez – Agustina Carando –
Samuel Davidson – Paloma Fernández Mira – Kenji Sagae
University of California, Davis / United States

Abstract – While research on second language (L2) tense-aspect acquisition has flourished, most studies have focused on lexical aspect as an explanatory variable (Bardovi-Harlig and Comajoan-Colomé 2020). However, the role of the features of first language (L1) production in L2 Spanish preterit-imperfect acquisition has never been tested before. Prior research has found that the frequency and distinctiveness of verb forms in corpora of L1 English production predict L2 English learners' tense-aspect production (Wulff *et al.* 2009). The present study aims to replicate these findings and test the predictions of hypotheses of L2 tense-aspect acquisition in another group of learners: English-dominant, instructed Spanish learners. Analyses were performed on longitudinal data from the *Corpus of Written Spanish of L2 and Heritage Speakers* (COWS-L2H; Yamada *et al.* 2020) and cross-sectional data from the *Corpus Escrito del Español L2* (CEDEL2; Lozano 2021). Results indicate that L1 verb frequency and distinctiveness predict learners' emergent use of the preterit and the imperfect.

Keywords – tense-aspect acquisition; Spanish as a second language; preterit and imperfect; learner corpus research

1. INTRODUCTION¹

The distinction between the preterit and imperfect is one of the most challenging Spanish grammatical concepts for learners whose first language (L1) does not mark aspectual differences through verbal morphology. For example, the difficulties experienced by L1 English learners of Spanish in accurately distinguishing between the perfective (preterit) and imperfective (imperfect) past have been repeatedly documented (cf. Bonilla 2013). In the context of Spanish language education in the United States, most instructors dedicate a substantial portion of their curriculum to explain and review these structures when teaching students how to tell stories in the past, share past experiences or talk about their weekends, vacations, etc. Given the difficulty and

¹The authors thank the COWS-L2H team members for their support of this project.
<https://ricl.aelinco.es/index.php/ricl/article/view/109>



importance of the structures, it is necessary to understand which factors influence learners' acquisition. While the field of second language acquisition (SLA) has thoroughly researched the acquisition of tense and aspect, including the acquisition of the preterit and the imperfect in Spanish, studies have focused on lexical aspect as an explanatory factor (cf. Bardovi-Harlig and Bergström 1996; Domínguez *et al.* 2013; González and Quintana Hernández 2018). However, a complete account of tense-aspect acquisition still awaits accurate description as it should take a broader range of predictors into account, such as form frequency, regularity, and saliency (Bayley 1994). In order to contribute to a more complete description of tense-aspect acquisition in second language (L2) Spanish, the present study investigates a predictor that has been under-researched in the field of preterit-imperfect acquisition: learners' mirroring of the distributional biases attested in L1 Spanish production.

After a review of the literature, Section 2 discusses the objectives and research questions in the study. Section 3 provides information on the research methodology. Section 4 and 5 constitute the core of the analysis and provide the results and their discussion. Finally, Section 6 offers a summary and some conclusions.

1.1. Tense and aspect

Tense and aspect describe the temporal positioning of an event and the interpretation or view of the event, respectively (Comrie 1985: 9). Tense, a deictic class, situates an event in relation to speech time, or the time at which the utterance is occurring. Aspect clarifies the way in which the event is viewed. The event may be viewed as bounded or as having a clear endpoint, in which case it would have perfective aspect (e.g. *María tocó el violín en el concierto* 'Maria played the violin in the concert'). In contrast, the event might be viewed as unbounded, or not having a clear endpoint, in which case it would have imperfective aspect (e.g. *María tocaba el violín todos los días* 'Maria used to play the violin every day'). Tense and aspect can be conveyed through verbal morphology as well as through other linguistic resources. Spanish has a rich verbal morphological system and encodes tense-aspect primarily through inflectional suffixation. Although the present study focuses on verbal tense-aspect marking, it is worth noting that the expression of tense-aspect also uses resources beyond the verb, such as the arguments of the predicate or adverbials (Verkuyl 1972; Bardovi-Harlig 2000; Bardovi-Harlig and Comajoan-Colomé 2020).

Prior studies on aspect acquisition have distinguished between the grammatical and lexical aspect of the verbal predicate. While ‘grammatical aspect’ refers to the encoding of aspectual meaning in the form of the verbal predicate (for instance, *querer* ‘want’ is marked with imperfective grammatical aspect through the imperfect form *quería* ‘I used to want’), ‘lexical aspect’ refers to the aspectual meaning that the inherent semantics of the verbal predicate conveys (Comrie 1976: 3). For example, the meaning of the verb *querer* ‘want’ conveys no clear input of energy from the subject, nor does it have a clear start and endpoint. In contrast, a verbal predicate like *summit a mountain* both requires energy to be dedicated to the action and has an inherent beginning and end (cf. Salaberry 2011: 187). These features of *querer* ‘want’ make it fall within the lexical aspectual category of ‘states’, while *summit a mountain* falls within the lexical aspectual category of ‘achievements’, as seen below. The following lexical aspect classification (Comrie 1976) has been used frequently in studies of aspect acquisition (cf. Andersen and Shirai 1994):

1. State: [- dynamic] [- punctual] [- telic] (e.g. *want*).
2. Activity: [+ dynamic] [- punctual] [- telic] (e.g. *ride a bicycle*).
3. Accomplishment: [+ dynamic] [- punctual] [+ telic] (e.g. *write an article*).
4. Achievement: [+ dynamic] [+ punctual] [+ telic] (e.g. *summit a mountain*).

With respect to the acquisition of tense and aspect, three main hypotheses have emerged, especially in consideration of how lexical aspect affects grammatical aspect marking. These hypotheses are described in what follows.

1.2. The Lexical Aspect Hypothesis

The Lexical Aspect Hypothesis (LAH) is based on Andersen’s (1991) seminal study on L1 English children’s naturalistic acquisition of Spanish as an L2. Andersen noticed that learners began marking verbs in the preterit before they did so in the imperfect. He found that the first verbs marked in the preterit denoted punctual events, such as *se partió* ‘something broke’, whereas the first verbs marked in the imperfect represented states, such as *tenía* ‘someone had’ (Andersen 1991: 314). He expanded on these findings to predict a general developmental sequence for tense-aspect acquisition, which constitutes the LAH (Andersen 2002).

The LAH also postulates that lexical aspect influences learners' choice of tense-aspect morphology most when learners are in the first stages of acquisition. Andersen and Shirai (1994) based this claim on a prototype model of grammatical and lexical aspect. According to this model, grammatical aspectual categories, such as the preterit, have more and less prototypical members. Lexical aspect is considered to be a primary factor in determining prototypicality. Following this analysis, telic predicates are prototypically associated with the preterit, and stative predicates are prototypically associated with the imperfect. The LAH predicts that prototypicality influences learners the most when they are beginners and states that more advanced learners will use the preterit and imperfect less prototypically. For example, learners are expected to use the preterit more preferentially with telic predicates when they are beginning to learn Spanish relative to subsequent periods of acquisition. However, this prediction has been contested (cf. Salaberry 1999, 2011).

Studies of L2 acquisition in different languages and contexts and with a variety of tasks have both supported and contradicted the LAH, as shown in Table 1.² Perhaps the most widely accepted tenet of the LAH is that lexical aspect plays a role in tense-aspect acquisition. While scholars have contested the specific route of tense-aspect development that the LAH proposes (Ayoun and Salaberry 2008) and the proposed effect of prototypicality on tense-aspect use at each proficiency level (Robison 1995), most scholars agree that lexical aspect affects grammatical aspect to some degree during L2 acquisition (Salaberry 2011). Based upon this consensus in the literature, the present study will not focus its research questions on testing lexical aspect as a factor but will rather focus on the impact of order of instruction and L1 distributional biases in the use of the preterit and imperfect.

Study	Focus	Participants	Task	Findings
Bardovi-Harlig and Bergström (1996)	LAH	L2 English (ESL) and L2 French (FFL) instructed learners	Written film retell task	Supported LAH
Salaberry (1999)	LAH	L2 Spanish instructed university students	Oral film retell task	Contradicted route predicted by LAH; impetus for DPTH
Salaberry (2002)	LAH, DPTH	L2 Spanish instructed university students	Written discourse-based cloze task	Supported DPTH

Table 1: Relevant studies on L2 tense-aspect acquisition

² For reviews on the literature, cf. Shirai 2004, Bonilla 2013, Bardovi-Harlig and Comajoan-Colomé 2020.

Study	Focus	Participants	Task	Findings
Wulff <i>et al.</i> (2009)	DBH/ distributional factors, lexical aspect	L2 English instructed university students and L1 English corpora	Oral interview task	Supported DBH and relevance of lexical aspect
Salaberry (2011)	LAH, DPTH	L2 Spanish instructed university students	Written discourse- based forced choice task	Supported DPTH
Collins <i>et al.</i> (2012)	Distributional factors, lexical aspect	Learner-directed speech corpus (for L2 English)	N/A	Supported DBH although did not explicitly test –English instructors– show distributional biases
Domínguez <i>et al.</i> (2013)	LAH, DPTH	L2 Spanish instructed high school and university students	Oral narration tasks and written sentence- context matching task	Supported DPTH
Thomas (2014)	LAH, distributional factors	L2 French instructed K-12 students, L1 French and learner- directed speech corpora	Oral conversations and narration tasks	Supported relevance of input frequency and lexical aspect to tense- aspect marking
González and Quintana Hernández (2018)	LAH, L1 influence	L2 Spanish instructed study abroad students	Written film retell task	Supported relevance of lexical aspect and L1 influence to tense-aspect marking
Tracy-Ventura and Cuesta Medina (2018)	DBH/ distributional factors	L1 Spanish corpora	N/A	Potentially supported DBH-L1 Spanish corpora show distributional biases, but did not consider L2 production
Daidone (2019)	DBH/ distributional factors	Learner-directed speech corpus and L1 Spanish corpora	N/A	Potentially supported DBH-Spanish instructors and L1 Spanish corpora show distributional biases, but did not consider L2 production
Izquierdo and Kihlstedt (2019)	Lexical aspect, L1 influence	L2 French instructed university students	Written film retell task	Supported relevance of lexical aspect and L1 influence to tense-aspect marking

Table 1: Continuation

1.3. The Default Past Tense Hypothesis

The Default Past Tense Hypothesis (DPTH) results from Salaberry's (1999, 2002) studies on the applicability of the LAH to instructed SLA (cf. Table 1). Salaberry (1999) examines L1 English college students' acquisition of the preterit-imperfect in L2 Spanish through a film retell task. In contrast to the prediction of the LAH that students initially rely on prototypical associations between lexical and grammatical aspect in

marking the preterit and imperfect, the study shows that students make more prototypical choices as their proficiency increases. Salaberry also finds that lexical aspect is not a significant factor in shaping preterit-imperfect production at the first stages of acquisition, as students mark the preterit on verbs of all lexical aspectual categories at this level. Salaberry explains this deviance from the LAH by claiming that the preterit is the default past tense marker for beginner instructed learners of Spanish, meaning that learners will use the preterit automatically or by default when they seek to mark the past in Spanish. Additionally, Salaberry (2002) notes that an instructional preference for teaching the preterit before the imperfect (i.e. an instructional effect), as well as cross-linguistic influence from the English simple past on preterit use,³ may contribute to the earlier emergence of the preterit. The DPTH thus postulates that L2 Spanish learners mark the past tense through the unmarked preterit form before they mark aspectual distinctions by introducing the imperfect into their linguistic repertoire (cf. Salaberry and Ayoun 2005; Salaberry 2008).

The DPTH has been most successful at predicting the development of instructed L2 learners, especially beginner learners whose L1 and L2 differ significantly in past tense-aspect marking, as is the case in L1 English learners of Spanish (cf. González and Quintana Hernández 2018; Bardovi-Harlig and Comajoan-Colomé 2020). In their study of L1 English acquisition of L2 Spanish, Domínguez *et al.* (2013) provide evidence supporting the DPTH. Their corpus and experimental data of spoken production and comprehension indicate that, in their study, beginner learners, who are year 10 high school students in the UK, mark verbal predicates of all lexical aspectual categories in the preterit. The beginner learners also show a preference for the preterit over the imperfect for every lexical aspectual category except states. Similarly, in a study of L2 writing, González and Quintana Hernández (2018) demonstrate that upper beginner L1 English learners of Spanish in a study abroad context show an overuse of the preterit. They attribute this pattern to cross-linguistic influence from the English simple past. As Salaberry (2002: 407) remarks, cross-linguistic influence may contribute to drive the initial preference for the preterit as predicted by the DPTH. The present study thus examines whether the DPTH's developmental sequence generalizes to a larger group of

³ The English simple past can be used both in perfective aspectual contexts and in imperfective aspectual contexts, as in *When I was a kid, I walked to school every day* which is the English equivalent of 'Cuando era niña, caminaba a la escuela cada día'. Thus, transfer from L1 English to L2 Spanish past tense-aspect marking may result from overextension of the preterit into imperfective aspectual contexts, based on analogy with the English simple past.

L1 English instructed Spanish language learners at beginner and intermediate proficiency levels.

1.4. The Distributional Bias Hypothesis

The Distributional Bias Hypothesis (DBH) does not solely consider verbs' lexical aspect but further explores how this verbal property may influence the distribution of verbs in the preterit and imperfect in the language of L1 speakers, and further considers how such a distribution may be replicated in L2 learners' production. The DBH (Andersen and Shirai 1994) is based on the finding that L1 speakers tend to use verbal predicates with a certain grammatical aspect category preferentially. Distributional biases occur in English as well as in Spanish (Tracy-Ventura 2007) and may occur in other languages. Andersen and Shirai (1994) claim that distributional biases are related to the effect of lexical aspect on grammatical aspect marking, as predicted by the LAH. For example, in the *EsPal Corpus*⁴ of L1 Spanish writing (Duchon *et al.* 2013), *tener* 'have' is used approximately two times more frequently in the imperfect than in the preterit. According to the DBH and the LAH, L1 Spanish speakers prefer to use *tener* 'have' with the imperfect because it is a stative verb. As the stative lexical aspect category is prototypically associated with the imperfective grammatical aspect, speakers produce *tener* 'have' with a bias toward the imperfect over the preterit.

The DBH predicts that learners will notice the preferential use of certain verbs with certain grammatical aspect categories. The memory capacity and data-driven learning ability of adult second language learners may facilitate learning based on distributional biases in L1 production (Shirai 2004: 109). Andersen and Shirai (1994) claim that when learners are exposed to L1 production, they often overgeneralize the link between a particular verb and its prototypical association with the preterit or the imperfect as found in L1 frequency biases. According to Shirai (2004), this phenomenon may explain the increase in learners' prototypical use of past tense-aspect marking as proficiency increases, as observed by Robinson (1995) and Salaberry (1999). Salaberry (2011) also advances that distributional biases may work in tandem with lexical aspect and other factors in order to determine learners' choice of grammatical aspect (cf. also Bardovi-Harlig and Comajoan-Colomé 2008). The DBH

⁴ <http://clic.ub.edu/corpus/es/espal>

thus goes beyond the LAH to explain tense-aspect acquisition in terms of interconnected semantic and cognitive factors. While the DBH has been proposed as a hypothesis, it has not been thoroughly tested in L2 tense-aspect literature and even less so in the context of L2 Spanish, as shown in Table 1.

Learners' emulation of distributional properties in L1 production has been considered as a pivotal factor in L2 tense-aspect acquisition, as exemplified by the postulation of the DBH (Andersen and Shirai 1994; Shirai 2004; Salaberry 2011; Ellis 2013; Thomas 2014). However, few studies have investigated the effect of this factor on tense-aspect acquisition at an empirical level (cf. Table 1). Wulff *et al.* (2009) are, to the best of our knowledge, the first to consider the role of features of L1 production in the acquisition of tense-aspect marking. The study compared beginner L2 English spoken production from a film retell task with two L1 English corpora of spoken production. Wulff *et al.* (2009) find that the frequency, distinctiveness, and prototypicality of the tense-aspect forms that are considered in the L1 English corpus predict the production of the forms in the L2 English corpus. The frequency of forms describes how often tokens occur in production. The 'distinctiveness' of forms refers to how closely a verb is associated with a particular tense-aspect category. Distinctiveness characterizes the frequency of a form contingent on its context of use and is therefore also referred to as 'contingent frequency' (cf. Wulff 2020: 177). The prototypicality of forms describes the extent to which a verb is a prototypical member of a tense-aspect category based on its inherent lexical aspect. Following Prototype Theory (Rosch and Mervis 1975), prototypical forms in any category hold the most integral features of a category and serve as a point of reference for category membership. For example, Wulff *et al.* (2009) report that the verb *run* is not only highly frequent in its progressive forms (e.g. *Someone is running*), but is distinctively associated with progressive aspect, and is a prototypical member of progressive aspect given its mid-range telicity score. The frequency, distinctiveness, and prototypicality that Wulff *et al.* find in the L1 English corpus is also mirrored in the L2 English corpus.

Several other studies have demonstrated the influence of L1 distributional properties on second language learning through a construction or usage-based grammar perspective. Ellis and Ferreira-Junior (2009), for instance, provide evidence that construction frequency, distinctiveness, and prototypicality may explain L2 English acquisition of verb-argument constructions, including the ditransitive construction (e.g.

Pat faxed Tom the picture). In alignment with findings from Wulff *et al.* (2009) and with constructionist theories (Goldberg 2003), Ellis and Ferreira-Junior (2009) find that the frequency of word types in a given construction follows a Zipfian distribution (Zipf 1935). In this type of distribution, the most frequent token occurs approximately two times more frequently than the second most frequent token, and three times more frequently than the third most frequent token (Wulff 2020: 178). Thus, the Zipfian distribution is characterized by an inverse relationship between token frequency and the token's relative order of frequency compared to the other tokens of the same construction. Ellis and Ferreira-Junior (2009) show that learners mirror the Zipfian distribution attested in L1 English production and employ the most distinctive and prototypical types of each verb-argument construction.

Investigations of L1 Spanish production have found distributional biases in the use of the preterit and imperfect (Tracy-Ventura and Cuesta Medina 2018; Daidone 2019). In oral texts from the *Corpus del Español* (Davies 2002), Tracy-Ventura and Cuesta Medina (2018) examine the frequency of past forms and show that both preterit and imperfect token frequency in the corpus follow a Zipfian distribution in which certain tokens represent a large percentage of the total preterit or imperfect tokens produced. A Distinctive Collexeme Analysis (henceforth DCA; Gries and Stefanowitsch 2004) shows that most of the tokens are also clearly associated with either the preterit or the imperfect. Tracy-Ventura and Cuesta Medina (2018) note that, in the texts analyzed, the verbs distinctly associated with the preterit are all telic, and those distinctly associated with the imperfect are all atelic. Their findings highlight that distributional biases occur in L1 Spanish production and that these biases relate to the lexical aspect of the verb, as proposed by Andersen and Shirai (1994).

Daidone (2019) catalogues the frequency of past forms in two corpora representing L1 Spanish and highly advanced L2 Spanish production, namely, learner-directed instructor speech from intermediate university classes, which is taken to represent classroom input, and oral texts from the *Corpus de Referencia del Español Actual* (CREA; *Real Academia Española*). A DCA demonstrates that the tokens in both corpora show biases toward the preterit or imperfect based on lexical aspect. The classroom input has greater biases toward the preterit, as the instructors rarely use imperfect forms; the preterit tokens represent 80 percent of the tokens analyzed. Daidone discusses instructors' preferential use of the preterit, as supporting Salaberry's

(2002) claim that the preterit may emerge before the imperfect in instructed learning because learners are exposed to a sufficient number of tokens in the preterit before they are exposed to a comparable amount of tokens in the imperfect. Daidone is the first to examine preterit-imperfect acquisition through a corpus of learner-directed classroom speech. Given that this corpus is not publicly accessible, studies on the role of L1 production properties in preterit-imperfect learning have primarily analyzed general L1 Spanish corpora (Tracy-Ventura and Cuesta Medina 2018). The present study adopts this approach with the understanding that future work will benefit from greater consideration of learner-directed classroom speech.

2. THE PRESENT STUDY

Bardovi-Harlig and Comajoan-Colomé (2020: 1128) describe LAH as “the single most influential hypothesis in second language acquisition (SLA) research regarding tense and aspect.” As seen in Table 1, lexical aspect has been an explanatory factor in studies of L2 tense-aspect acquisition for more than two decades. In contrast, very few studies have addressed learners’ mirroring of the frequency biases in L1 Spanish production as a factor while investigating learners’ production of past tense-aspect, apart from a few isolated references here and there (cf. Wulff *et al.* 2009; Thomas 2014), and no studies to date have examined this factor in L2 Spanish preterit-imperfect acquisition. In order to model the complex process that L2 tense-aspect development implies, the field must examine more thoroughly the multitude of factors that influence this process, in addition to lexical aspect.

Given the recent advances in methods for Learner Corpus Research (LCR), studying distributional features of L1 Spanish production as an explanatory variable is currently more feasible and effective than it was when the LAH was proposed. To expand research on a multifactor account of tense-aspect acquisition, the present study considers two corpora of Spanish writing. As these corpora have never been considered in investigations of tense-aspect learning, the study tests the generalizability of findings in prior studies. Prior studies have also favored cloze tasks and film retell tasks (cf. Table 1), which limit the range of verb types that learners produce when compared to open-ended production tasks. This study examines production in open-ended writing tasks that are not scaffolded for preterit-imperfect elicitation in order to confirm that the results attested in more structured elicitation tasks apply generally. Finally, the

investigation deals with calls within LCR for greater linguistic description in addition to statistical testing (Larsson *et al.* 2022). Crucially, the study investigates preterit-imperfect development by employing infrequently used task types and following recommendations to highlight linguistic phenomena in learner texts.

The present study aims to nuance our current understanding of preterit-imperfect acquisition by expanding beyond the LAH to consider an understudied factor in tense-aspect acquisition: learners' emulation of the distributional biases in L1 Spanish production. The following research questions assess the predictions of each of the hypotheses (LAH, DPTH and DBH):

1. Does the LAH and DPTH's prediction⁵ that instructed Spanish learners will produce the preterit before the imperfect correspond with the developmental trajectory of the preterit and the imperfect observed in the learner sample?
2. How closely is the token frequency of past forms in learners' production associated with the token frequency in L1 Spanish production?
3. How closely does the contingent frequency (distinctiveness) of past forms in learners' production reflect the contingent frequency in L1 Spanish production?

3. METHODS

3.1. Design

In order to capture preterit-imperfect acquisition over time from a large and diverse sample of participants, the study is divided into two parts. First, a longitudinal study is conducted with beginner students from the University of California, Davis, a large, public university in the U.S. These students contributed writing samples to the corpus over the course of three academic terms, which constituted one academic year in total. The longitudinal study provides evidence of student development over time, which is contextualized through the curriculum of the language program. Learner writing samples from the longitudinal group exemplify how the preterit and the imperfect emerge in interaction with the essay genre. Secondly, a cross-sectional study has been

⁵ Certainly, the LAH and DPTH make different predictions about the route of emergence of the preterit-imperfect and attribute these routes to different factors (e.g. lexical aspect vs. a default past tense form). However, the hypotheses concur in their prediction that the preterit will emerge before the imperfect for L2 learners of Spanish. Therefore, the first research question tests this prediction in both hypotheses.

conducted with 1) a learner corpus representing a large and varied sample of students, and 2) a corresponding L1 Spanish reference corpus. The cross-sectional analysis has allowed for comparison between L1 and L2 use of the preterit and imperfect.

3.2. Corpora

Current L2 Spanish corpora offer a variety of types of data that facilitate the study of L2 grammatical development. This study takes advantage of the unique characteristics of the two largest corpora of written L2 Spanish: 1) the *Corpus Escrito del Español L2* (CEDEL2; cf. Lozano 2009, 2021; Lozano and Mendikoetxea 2013) and 2) the *Corpus of Written Spanish of L2 and Heritage Speakers* (COWS-L2H; Yamada *et al.* 2020). CEDEL2 features a large L1 Spanish reference corpus, which has been taken as a representation of L1 Spanish writing. The written, not the spoken, L1 Spanish data has been used in order to control for modality and task effects, as the L1 and L2 data have been elicited from identical tasks. COWS-L2H offers a longitudinal student sample from one university setting, namely, the University of California, Davis, whereas CEDEL2 provides a cross-sectional learner sample from a wide variety of instructional settings (e.g. Denison University, Georgia State University, Pennsylvania State University). In order to control for the variables of textual genre and text length across corpora, only descriptive and narrative essays between 50 and 500 words in length have been analyzed. The use of tasks that were not explicitly intended to elicit preterit-imperfect production have allowed the study to capture L1 and L2 Spanish production in a more ecologically valid manner, without the greater potential for priming effects that may occur in more structured elicitation tasks, such as interviews (Izquierdo and Kihlstedt 2019).

3.2.1. CEDEL2

CEDEL2 includes essays written by L2 Spanish learners and L1 Spanish speakers representing several different varieties of Spanish. The tasks consist of unmonitored online writing assignments without time constraints. The volunteer sample of participants chooses to respond to one of 14 prompts that are proposed on the project's website. The cross-sectional analysis of CEDEL2 includes 820 L1 Spanish essays and 611 L2 Spanish essays, all of which are written by different participants. The mean

essay length is 231 words in the L1 Spanish group and 206 words in the L2 Spanish group. Only learners who reported their L1 as English and their age as between 17 and 26 years are included. Learners range from A1 to B2 on the scale of the Common European Framework of Reference for Languages (CEFR; Council of Europe 2001), as displayed in Table 2. Proficiency level is determined by students' scores on the University of Wisconsin (1998) college-level placement test, which they complete at the time of data collection.

Proficiency		CEDEL2	
Proficiency level	Proficiency level CEFR	Number of participants	Number of essays
Lower beginner	A1	29	29
Upper beginner	A2	186	186
Lower intermediate	B1	192	192
Upper intermediate	B2	204	204

Table 2: Participants, essays, and proficiency levels in the cross-sectional study

3.2.2. COWS-L2H

COWS-L2H includes essays written by students enrolled in Spanish classes at a large, public university in the United States. Students complete a Web-based Computer Placement Exam (WebCAPE 2.0) in order to be placed into a class. The placement scores corresponding to the course levels of the participants fall between A1 and A2 proficiency on the CEFR scale (Yamada *et al.* 2020; Fernández-Mira *et al.* 2021). Similar to CEDEL2, writing tasks are unmonitored, completed online, and without time constraints. All students respond to a descriptive prompt in the fourth week of each academic term and a narrative prompt in the eighth week of the ten week-long academic term. Participants who volunteer to participate in the corpus study, which is separate from their normal coursework, may do so during multiple academic terms and are compensated with course extra credit. The student sample does not constitute an intact class, as the students are not all enrolled in the same Spanish class. Several students participate repeatedly in different terms throughout the four years of data collection, thus providing a relatively large set of longitudinal data.

The longitudinal research in the study is conducted using the written samples of eight students who participate six times in the first three academic terms of the Spanish program (cf. course levels SPA 1, SPA 2, and SPA 3 in Table 3). The first-year Spanish

program (SPA 1-3) focuses on the development of basic communicative skills. In the CEFR scale, SPA 1 corresponds to the A1 proficiency level, SPA 2 to A1+, whereas SPA 3 relates to A2 (cf. Table 3; Fernández-Mira *et al.* 2021). The 48 essays analyzed have a mean length of 218 words.

Course level	Proficiency level	Proficiency level CEFR	Number of essays	Number of participants
SPA 1	Lower beginner	A1	16	8
SPA 2	Lower beginner	A1+	16	8
SPA 3	Upper beginner	A2	16	8

Table 3: Course, proficiency, essays, and participants in the longitudinal study

Only participants with no prior experience learning Spanish and who report their L1 as English do not produce the preterit-imperfect before they are taught the structures in SPA 2. As these students are real beginners, they do not a priori have classroom-based knowledge of the preterit-imperfect in SPA 1. Therefore, we assume that students who use the preterit-imperfect in SPA 1 have either 1) some exposure to Spanish at home or in their community or 2) do not follow the task's instructions which do not allow them to consult outside resources like online translators. For this reason, only true beginners who do not produce the preterit-imperfect in SPA 1 have been included. In order to demonstrate the emergence of the preterit and imperfect in students' writing over time, the study highlights samples from participants' essays. These samples facilitate the description of the linguistic features as they appeared in students' writing, which is an essential element of corpus analysis (Larsson *et al.* 2022).

3.3. Procedure

The corpora are tokenized and tagged for part-of-speech using *FreeLing 4.2* (cf. Padró *et al.* 2010; Padró and Stanilovsky 2012). *FreeLing* tags verbs for tense, aspect, and mood, among other features, with 97 percent accuracy (Padró and Stanilovsky 2012). For each essay in CEDEL2 and COWS-L2H, the tokens tagged with *FreeLing* as verbs in the preterit or imperfect indicative have been collected in their token and lemma forms in *Python 3.9*.⁶ These forms have been analyzed in terms of their token, lemma, and contingent frequency. Contingent frequency has been measured using a DCA analysis (Gries and Stefanowitsch 2004). The DCA measures the strength of association

⁶ <https://www.python.org/downloads/>

between a verb and the preterit, or a verb and the imperfect, based on its frequency of use in the preterit/imperfect relative to its total frequency of use.

4. RESULTS

4.1. Longitudinal study

According to the LAH and the DPTH, the preterit should emerge before the imperfect among instructed L1 English learners of Spanish. The learners in this longitudinal study do not report prior experience learning Spanish; hence, their developmental trajectory is based on instruction at the university. The average of these students' preterit and imperfect token production at each data collection time is considered to gauge development over the course of three academic terms. The students' mean use (per 100 words) in each essay is reported in order to account for differences in text lengths. Variance is measured through the standard deviation, as shown in Table 4, and 95 percent confidence interval of the mean, as seen in the error bars of Figure 1.

Data collection time	Mean preterit tokens per 100 words (SD)	Mean imperfect tokens per 100 words (SD)
SPA 1 midpoint	0 (0)	0 (0)
SPA 1 endpoint	0 (0)	0 (0)
SPA 2 midpoint	3.42 (3.32)	0.25 (1.29)
SPA 2 endpoint	3.64 (2.54)	2.64 (2.04)
SPA 3 midpoint	1.04 (1.61)	0.65 (1.41)
SPA 3 endpoint	2.92 (2.07)	2.98 (2.49)

Table 4: Preterit and imperfect number of tokens per 100 words

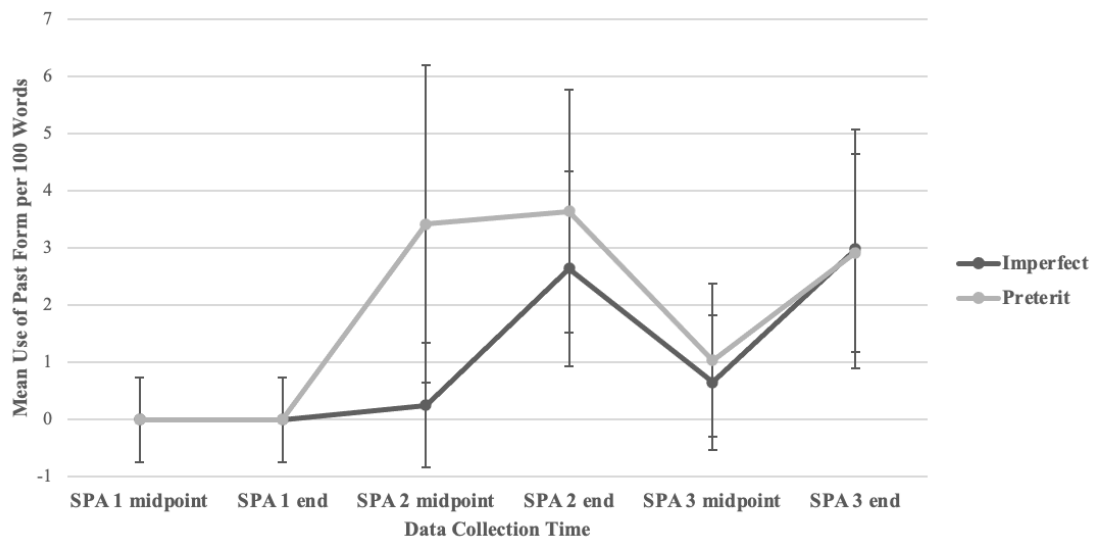


Figure 1: Longitudinal development of preterit and imperfect token production

The instructional effect of students learning the preterit at the beginning of SPA 2 is clearly seen, as the students make a more frequent use of the preterit on average: 3.42 times per 100 words at the midpoint of SPA 2. Meanwhile, the students do not frequently use the imperfect at the midpoint of SPA 2: mean use of 0.25 (cf. Table 4). The following excerpts, (1)–(2), exemplify how these students start to use the preterit, often in contexts where the imperfect would be more acceptable. For example, in (1), the student uses the preterit (e.g. *comimos* ‘we ate’, *compró* ‘he/she bought’) to describe habitual actions in the past, which would typically be marked in the imperfect:

- (1) En su coche [de mi padre], nosotros **cantamos** PRET muchos canciones y **reímos** PRET porque mi padre **estuvo** PRET fuerte cuando él **cantó** PRET. Nosotros **comimos** PRET pan tostado todas las mañanas y los sábados, **montamos** PRET nuestras bicicletas y **fuimos** PRET al parque. Mi padre siempre me **compró** PRET helado en el parque. (Female, 21, prompt: describe a special person)

‘In his car [of my father], we **sang** many songs and **laughed** because my father **was** loud when he **sang**. We **ate** toast every morning and on Saturdays, we **got on** our bikes and **went** to the park. My father always **bought** me ice cream at the park.’

- (2) El año pasado, ella [Beyonce] **hizo** PRET Lemonade, los discos compactos. Ella **tuvo** PRET gemelos el año pasado también, por lo que ella **es** PRES trabajadora. Ella **cantó** PRES en Coachella. **Estuve** PRET muy celosa de mis amigos que la **vieron** PRET, pero **voy a** FUT verla cantar en Septiembre en Santa Clara. (Female, 20, prompt: describe a famous person)

‘Last year, she [Beyonce] **made** Lemonade, the albums. She **had** twins last year too, she is so hardworking. She **sang** in Coachella. I **was** very jealous of my friends who saw her, but I **am going to** see her sing in September in Santa Clara.’

The instruction of the imperfect in the latter half of SPA 2 corresponds with an increase in the average usage of the imperfect to 2.64 times at the end of SPA 2. The preterit is still more frequently used than the imperfect at this level, on average 3.64 times. As can be noticed in (3)–(4), the students begin to use the imperfect for long stretches of text, often alternating with the preterit. The more recent instruction of the imperfect even leads to an overuse of the imperfect when the preterit is necessary as, for instance, with the use of *veíamos* (first person plural imperfect past form of *ver* ‘see’) in (3).

- (3) Una día, mi y mi amiga **caminábamos** IMP en mi ciudad. La día **estaba** IMP muy soleado. Cuando nosotros **íbamos** IMP a la tienda, nosotros **veíamos** IMP nuestra otra amiga. ¡Hola ellas!, nuestra amiga **habló** PRET. Nosotros **hablamos** PRET a juntos por dos horas. Después la tienda, nosotros **íbamos** IMP a él café... (Female, 18, prompt: tell a terrible story)

‘One day, me and my friend were *walking* in my city. The day *was* very sunny. When we *were going to* the store, we were *seeing* our other friend. Hi [girls]! our friend *spoke*. We *spoke* together for two hours. After the store, we *were going to* the café...’

- (4) Pero la mejor parte de estas vacaciones *era* IMP la selva. **Condujimos** PRET a la selva temprano en el mañana. *Estaba* IMP muy lejos. **Caminamos** PRET por la selva todo el día. *Estaba* IMP muy caliente y húmedo. **Vimos** PRET muchos animales como serpientes, arañas y pájaros. (Female, 20, prompt: narrate your perfect vacation)

‘But the best part of this vacation *was* the rainforest. We **drove** to the rainforest early in the morning. It *was* very far away. We **walked** through the rainforest all day. It *was* really hot and humid. We **saw** many animals like snakes, spiders, and birds.’

An effect of textual genre on past tense use is especially visible in SPA 3. Here students respond to descriptive prompts at the midpoint, and narrative prompts at the end of each term. As evidenced by the difference between the midpoint and end of SPA 3, students produce more past tense tokens in the narrative genre relative to the descriptive genre. Interestingly, on average, students produce similar amounts of imperfect (2.98) and preterit tokens (2.92) at the end of SPA 3. This demonstrates that the preterit is no longer the default by the end of SPA 3.

In sum, in this sample students do not produce the preterit and imperfect until the structures are taught. When each structure is taught, students greatly increase their use of the structure, sometimes overextending it to encompass more past functions than what is grammatically acceptable. This result provides additional evidence to support the role of the order of instruction in preterit-imperfect production (Salaberry 2002).

4.2. Cross-sectional study

4.2.1. Emergence of the past

The predictions of the LAH and DPTH have also been tested in the CEDEL2 cross-sectional data to assess the generalizability of the hypotheses across learner groups. Figure 2 visualizes the mean preterit-imperfect usage per 100 words in each essay with error bars denoting the 95 percent confidence interval. As seen in Figure 2 and Table 5, corpus data show a preference for the preterit at all proficiency levels under investigation. At the A1 level, on average, students produce 1.08 preterit and 0.78 imperfect tokens per 100 words. The mean use increases for the preterit but decreases

for the imperfect at the A2 level. This is likely due to students writing longer essays at the A2 level than at the A1 level, without a corresponding increase in imperfect use. From the A2 to the B1 level, the mean number of imperfect tokens increases slightly to almost 1 and the mean number of preterit tokens jumps to almost 3. Mean use of the imperfect shows a dramatic increase at the B2 level: up to 1.65 tokens. Students produce preterit tokens on average 3.11 times at the B2 level.

Proficiency level	Mean preterit tokens per 100 words (SD)	Mean imperfect tokens per 100 words (SD)
A1	1.08 (2.20)	0.78 (2.36)
A2	1.61 (2.78)	0.58 (1.36)
B1	2.87 (3.35)	0.99 (1.85)
B2	3.11 (3.25)	1.65 (2.18)

Table 5: Mean number of preterit-imperfect tokens per essay: A cross-sectional study

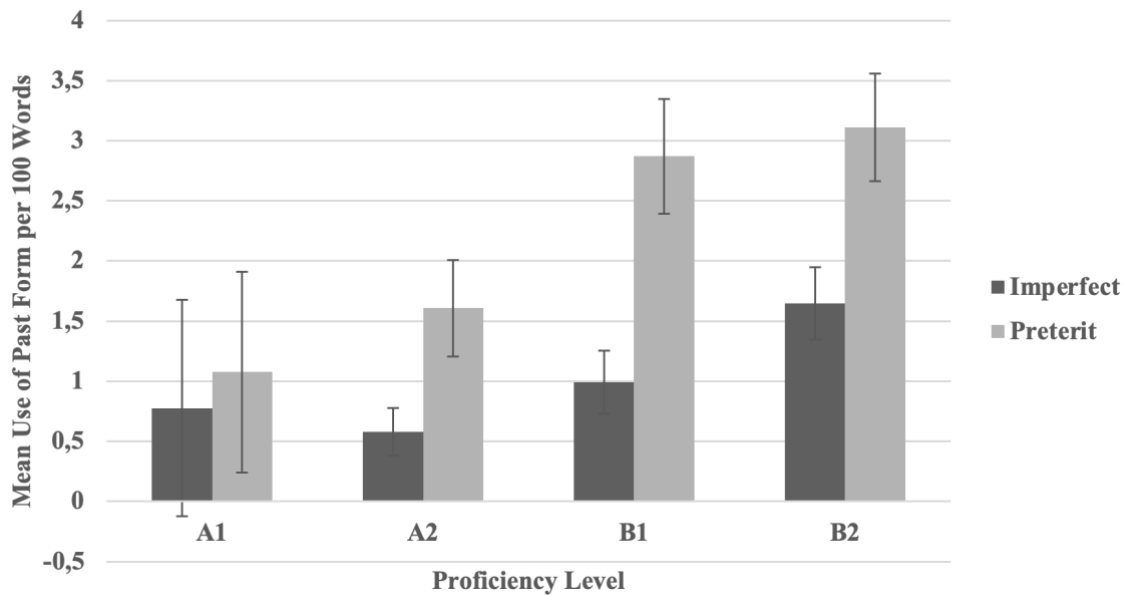


Figure 2: Cross-sectional development of preterit and imperfect token production

4.2.2. The Distributional Bias Hypothesis (DBH)

Having confirmed that the LAH and the DPTH accurately predict the prevalence of the preterit in early stages of L2 development, the question stands: are all verbs equally prone to being used in the preterit or the imperfect? Or, as proposed by the DBH, are L1 biases in the use of specific verbs in specific tenses also reflected in L2 writing? To answer our second research question, the study investigates the distributional biases in L1 Spanish and how these biases are reflected in L2 Spanish production, comparing the

CEDEL2 learner sample with the CEDEL2 L1 sample. Both groups complete the same writing task, and the comparison of essays from the same task control for the effects of differences in textual genre on writing.

4.2.3. Token Frequency Distribution

Prior studies on construction learning have demonstrated that the frequency of tokens in a construction typically follows a Zipfian distribution (Wulff 2020) which is characterized by an inverse relation between token frequency and rank order of frequency among the tokens. The present study considers the 27 most frequent verbs in the preterit and imperfect in the L1 and L2 CEDEL2 corpora to determine whether their frequencies fit a Zipfian distribution.

As seen in Figure 3, the CEDEL2 L1 Spanish sample follows a Zipfian distribution. *Ser* ‘be’, which has been the verb most frequently attested in the preterit, is approximately 2.5 times more frequent than *ir* ‘go’, the second most frequent verb in the preterit. Following *ir* ‘go’ and the third most frequent preterit verb, *encontrar* ‘find’, the frequency of preterit verbs decreases gradually. In the imperfect, *haber* ‘have’, *ser* ‘be’, and *estar* ‘be’ are the three most frequent verbs. These verbs are approximately two times more frequent than *tener* ‘have’, the fourth most frequent verb.

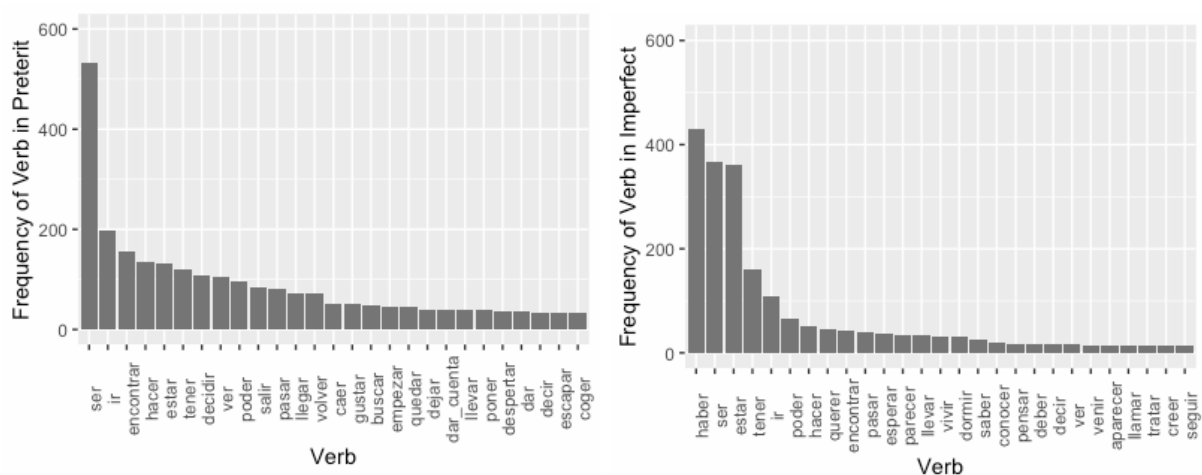


Figure 3: Frequency distribution of verbs in preterit (left) and in imperfect (right) in CEDEL2-L1

As seen in Figure 4, the CEDEL2 L2 Spanish sample follows a slightly more distinctive Zipfian distribution than the L1 Spanish sample. *Ser* ‘be’ and *ir* ‘go’ are still the two most frequent verbs in the preterit, and *ir* ‘go’ is approximately three times more

frequent than *ver* ‘see’, the third most frequent preterit verb. In the imperfect, *ser* ‘be’ is the most frequent verb and is approximately 1.5 times more frequent than *estar* ‘be’, the second most frequent verb. The decrease in the frequency of both preterit and imperfect verbs is slightly less marked in the L2 corpus when compared to the L1 corpus.

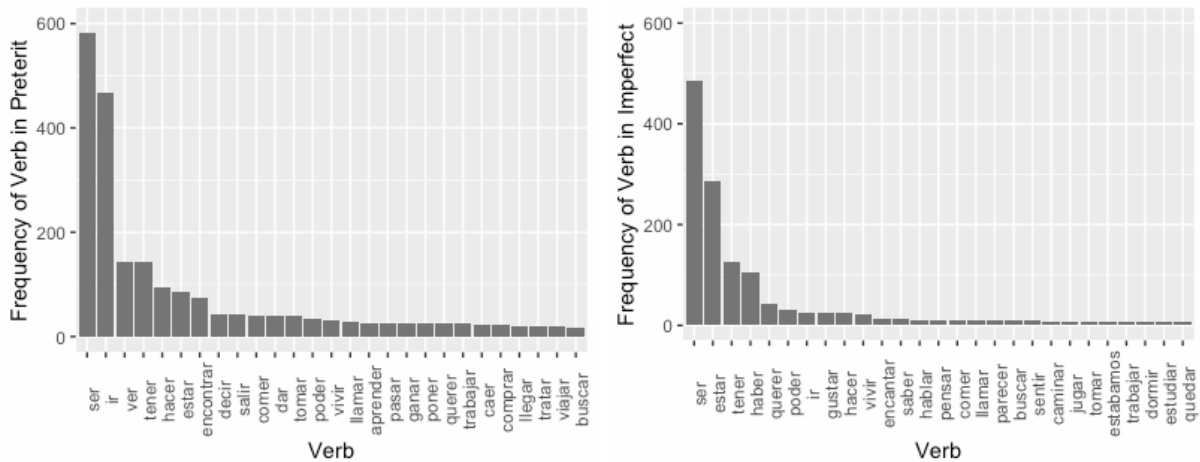


Figure 4: Frequency distribution of verbs in preterit (left) and imperfect (right) in CEDEL2-L2

4.2.4. Association between L1 and L2 production

A linear regression (*stats* 4.0.2 in R)⁷ has been conducted to assess the relation between L1 Spanish speakers’ and learners’ frequency of past tokens. As may be observed in Figure 5, results show an R^2 value of 0.78 ($r = 0.88$, $p < 2.2e-16$) for the relation between L1 and L2 frequency regarding the 300 verbs in the preterit. The relation for the 129 verbs in the imperfect shows an R^2 value of 0.70 ($r = 0.83$, $p < 2.2e-16$), denoting a slightly stronger correlation in the preterit than in the imperfect. The main effect of L1 frequency is marginally weaker in the imperfect ($t = 17.14$) than in the preterit ($t = 32.36$). However, tokens that are highly frequent in the L1 corpus, such as *fue* ‘someone went/was’ in the preterit and *tenía* ‘someone had’ in the imperfect, have almost equivalently high frequencies in L1 corpus as in the L2 corpus. In sum, the correlation between L1 and L2 token frequency is large for both the preterit and the imperfect.

⁷ <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html>

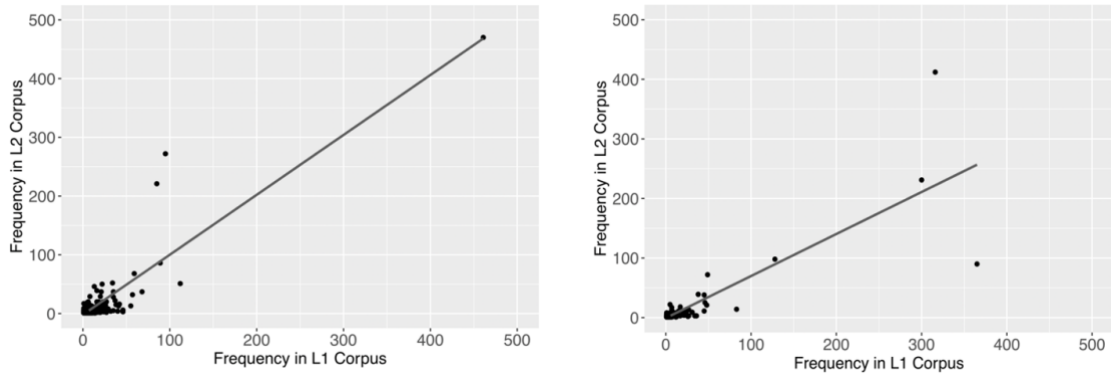


Figure 5: Relation between frequency of tokens in the preterit (left) and the imperfect (right) in CEDEL2 L1 and L2

4.2.5. Contingent frequency

The contingent frequency analysis considers the association between verbs and the preterit or imperfect. We have conducted a DCA (Gries and Stefanowitsch 2004) which yields results for the association between lemmas and the constructions with which they occur, such as the preterit and the imperfect. The *R-script coll.analysis* (Gries 2014) has been used to measure the association strength through binomial tests. The script calculates the observed and expected frequency of lemmas with each construction and returns the log likelihood, labeled as collostructional strength. Collostructional strength values above 1.3 indicate a significant p -value at the 95 percent confidence level ($p < 0.05$). Verbs with the largest collostructional strength values are considered the most distinctive verbs in the preterit or imperfect.

As seen in Tables 6 and 7, L1 and L2 Spanish speakers show distinctive associations at the 95 percent confidence level for at least ten verbs in the preterit and the imperfect. The distinctive verbs for the L1 and L2 speakers do not entirely overlap; 40 percent are the same in the preterit, and 50 percent are the same in the imperfect.

Preterit					Imperfect			
Verb rank	Verb	Imperfect token frequency	Preterit token frequency	Collostruction strength	Verb	Imperfect token frequency	Preterit token frequency	Collostruction strength
1	<i>Decidir</i> 'decide'	1	109	21.35	<i>Haber</i> 'have'	431	18	161.87
2	<i>Llegar</i> 'arrive'	6	73	9.32	<i>Estar</i> 'be'	362	132	60.03
3	<i>Ver</i> 'see'	16	105	9.06	<i>Tener</i> 'have'	160	120	10.38
4	<i>Volver</i> 'return'	7	73	8.67	<i>Querer</i> 'want'	46	11	10.28
5	<i>Salir</i> 'go out'	11	85	8.39	<i>Esperar</i> 'wait'	36	7	9.01
6	<i>Darse cuenta</i> 'realize'	0	38	7.96	<i>Dormir</i> 'sleep'	30	6	7.49
7	<i>Caer</i> 'fall'	3	52	7.69	<i>Saber</i> 'know'	25	5	6.32
8	<i>Coger</i> 'catch'	0	33	6.91	<i>Tratar</i> 'try'	14	1	4.88
9	<i>Encontrar</i> 'find'	44	157	6.37	<i>Deber</i> 'must'	17	3	4.65
10	<i>Escapar</i> 'escape'	1	34	5.97	<i>Necesitar</i> 'need'	10	0	4.19

Table 6: DCA Results for CEDEL2-L1

Preterit					Imperfect			
Verb rank	Verb	Imperfect token frequency	Preterit token frequency	Collostruction strength	Verb	Imperfect token frequency	Preterit token frequency	Collostruction strength
1	<i>Ir</i> 'go'	25	467	51.64	<i>Estar</i> 'be'	287	87	77.52
2	<i>Ver</i> 'see'	2	144	21.16	<i>Haber</i> 'have'	105	11	40.38
3	<i>Encontrar</i> 'find'	5	75	7.53	<i>Ser</i> 'be'	486	582	26.72
4	<i>Salir</i> 'Go out'	2	44	5.24	<i>Querer</i> 'want'	44	25	7.48
5	<i>Decir</i> 'say/tell'	3	44	4.52	<i>Tener</i> 'have'	125	143	7.04
6	<i>Aprender</i> 'learn'	0	26	4.31	<i>Gustar</i> 'like'	24	11	5.13
7	<i>Caer</i> 'fall'	0	23	3.81	<i>Encantar</i> 'love'	14	3	4.63
8	<i>Dar</i> 'give'	4	39	3.27	<i>Saber</i> 'know'	14	7	3
9	<i>Ganar</i> 'win'	1	25	3.19	<i>Poder</i> 'can'	31	35	2.2
10	<i>Hacer</i> 'do'	24	93	2.33	<i>Sentir</i> 'feel'	9	5	1.92

Table 7: DCA results for CEDEL2-L2

5. DISCUSSION

The longitudinal and cross-sectional data in the study demonstrate that learners generally produce the preterit more frequently than the imperfect. In the study, learners show an earlier increase in usage, which is characteristic of the emergence of a tense-aspect form, for the preterit rather than for the imperfect. The longitudinal data from COWS-L2H has also exemplified the influence of textual genre on preterit-imperfect production. Narrative essays, which students wrote at the end of each academic term, consistently elicit more preterit and imperfect tokens than the descriptive essays written at the midpoint of the term. This is in line with prior findings on the effect of genre on tense-aspect production (Bardovi-Harlig and Comajoan-Colomé 2020). Future studies would benefit from keeping textual genre constant at all stages in the longitudinal study. Nonetheless, the data from the end of each term in the longitudinal study and the data of mixed textual genre in the cross-sectional study clearly show that the emergence of the preterit precedes that of the imperfect.

Concerning the three research questions in the study, the first of them has tested the LAH and DPTH and the result confirms the predictions of the LAH and DPTH, namely, that the preterit generally emerges before the imperfect in students' writing in the COWS-L2H longitudinal study. The DCA also reveals that certain verbs are distinctly associated with the preterit and the imperfect in both corpora. Based on this finding, it seems clear that lexical aspect plays a role in the distributional biases. The verbs retrieved in the study which are highly distinctive of the preterit, such as *encontrar* 'find', *salir* 'go out', and *caer* 'fall', are primarily telic verbs. The verbs that are highly distinctive of the imperfect, such as *haber* 'have', *estar* 'be', and *tener* 'have', are primarily stative verbs. The contingent frequency analysis therefore provides indirect evidence of the role of prototypicality, as predicted by the LAH, in learners' use of the preterit and the imperfect.

The results are in line with Daidone's (2019) and Salaberry's (2002) conclusions in that the sequential instruction of the preterit and imperfect, as well as cross-linguistic influence from English, may contribute to an early preference for the preterit. As attested in the text samples, the longitudinal participants greatly increase their production of preterit and imperfect forms during the period directly following the instruction of each construction. The temporal alignment between the increase in the use

of the preterit and the introduction of the constructions in the curriculum highlights the relevance of explicit instruction in the acquisition of these constructions.

The second and third research questions have tested the DBH. In line with prior studies on the distribution of past tense-aspect forms in L1 corpora (Tracy-Ventura and Cuesta Medina 2018), the frequency of preterit and imperfect verbs in the L1 corpus follows a Zipfian distribution. The L2 data seems to fit the Zipfian distribution more closely than the L1 data; the most frequent verbs in each construction in the L2 data constitute a larger portion of the total preterit and imperfect verbs produced. Learners' limited lexicon is likely responsible for this skewed distribution. While learners may use highly frequent verbs like *ser* 'be' at the same, or greater, frequency relative to L1 speakers, they do not use less frequent verbs as often as L1 speakers. Overall, the distributional analysis proves that both L1 and L2 speakers show distributional biases in their production of the preterit and the imperfect.

Research question 2 has evaluated L1 token frequency as a factor in L2 preterit-imperfect production. The strong main effects in both the preterit ($t = 32.36$) and the imperfect ($t = 17.14$) provide evidence that learners are clearly attuned to token frequency in L1 production. Learners' exposure to L1 speakers using certain verbs in the preterit and imperfect plays a role in the relative frequency with which they mark certain verbs in the preterit and imperfect in their own writing. In our study, L1 token frequency has been a stronger predictor of L2 token frequency in the preterit ($R^2 = 0.78$) than the imperfect ($R^2 = 0.70$), which is likely due to the greater irregularity of preterit morphology. As there are more morphologically irregular forms in the preterit, learners are likely to acquire the preterit in a more item-based manner than the imperfect (MacWhinney 2016). This may result in learners more closely mirroring L1 Spanish frequency in the preterit than in the imperfect. While the association between L1 and L2 frequency demonstrates that L1 token frequency is a pivotal factor in shaping L2 production, explaining 70–78 percent of the variance in L2 preterit-imperfect token frequency, other factors likely account for the remaining variance. These factors may include distinctiveness and prototypicality as well as form regularity, saliency, and explicit instruction (Salaberry and Ayoun 2005).

Research question 3 has evaluated L1 contingent frequency (distinctiveness) as a factor in L2 preterit-imperfect production. In the study, several verbs have been distinctly associated with the preterit or the imperfect in both corpora, which provides

further evidence for the existence of a distributional bias. Only 40–50 percent of the ten most distinctive verbs were the same in the L1 and L2 corpora. Many of these verbs do not have a significant association with the preterit or the imperfect in the other corpus. For example, *esperar* ‘wait’, *dormir* ‘sleep’, *deber* ‘must’, and *necesitar* ‘need’ were distinctly associated with the imperfect in the L1 but not in the L2 corpus. Learners’ limited lexicon and their uncertainty about the conditions for the use of the preterit or imperfect may explain this difference. While the L1 group produces *esperar* ‘wait’ 43 times in the preterit or imperfect, the L2 group produces the verb only five times. The fact that the L2 learners produce the verb so few times in the past may indicate that the learners lack familiarity with the verb, its contexts of use in the past, and/or its grammatical marking in the past. It seems clear that both L1 and L2 Spanish speakers are attentive to contingent frequency in their use of the preterit and imperfect. The differences between the groups in their distinctive associations reflect additional factors in acquisition, including vocabulary development.

In sum, all three hypotheses (LAH, DBH, and DPTH) of past tense-aspect acquisition accurately predict facets of the L2 Spanish learners’ development and production of the preterit and imperfect. The LAH and DPTH highlight the earlier emergence of the preterit, which may be caused by the order of instruction of the structures and cross-linguistic influence from English, as well as by prototypical lexical-grammatical aspect associations. The DBH explains the role of learners’ mirroring of properties of L1 Spanish production, including the distributional biases that result in certain verbs having stronger contingent frequencies with a past tense-aspect form than others. The contingent frequencies in the L1 and L2 corpora show differences in part because learners’ lexicons are limited when compared to L1 Spanish writers. As learners acquire more lexical items, it is anticipated that their frequency distributions of preterit-imperfect marking will further approach L1 writers’ distributions. Given that this is the first study to empirically test the DBH for L2 Spanish, the strength of L1 frequency as a predictor of L2 preterit-imperfect production demonstrates a need for greater consideration of the DBH in explanations of past tense-aspect learning.

6. SUMMARY AND CONCLUSIONS

In research on tense-aspect acquisition, there is a need for studies in a wider variety of typologically different languages that consider factors beyond lexical aspect, such as

learners' emulation of distributional properties of L1 production, in order to achieve a more conclusive picture on the acquisition of these constructions. The present study is a first step in that direction and fills a gap in the literature by evaluating the predictions of three hypotheses of tense-aspect acquisition for L2 Spanish learning of the preterit and the imperfect. The predictions of the DPTH and LAH about learners' early preference for the preterit generalized to longitudinal and cross-sectional learner data in two Spanish learner corpora. In accordance with the DBH, the token frequency in the preterit and the imperfect followed a Zipfian distribution in both groups, indicating that both use certain verbs with a bias toward the preterit or the imperfect. In our data, learners mirror the token and contingent frequency of verbs in the L1 corpus, providing evidence that learning based on the properties of frequency and distinctiveness in L1 Spanish production occurs for the L2 Spanish acquisition of the preterit and the imperfect. This finding proves that Wulff *et al.*'s (2009) conclusion for L2 English oral production generalizes to L2 Spanish writing. Our study is the first to establish that L1 Spanish token and contingent frequency are strong predictors of L2 Spanish preterit-imperfect marking. Future research would benefit from examining accuracy of production and learners' mirroring of other distributional properties in L1 production, including regularity and phonological saliency.

REFERENCES

- Andersen, Roger W. 1991. Developmental sequences: The emergence of aspect marking in second language acquisition. In Thom Huebner and Charles Ferguson eds. *Crosscurrents in Second Language Acquisition and Linguistic Theories*. Amsterdam: John Benjamins, 305–324.
- Andersen, Roger W. 2002. The dimensions of 'pastness'. In M. Rafael Salaberry and Yasuhiro Shirai eds. *The L2 Acquisition of Tense-Aspect Morphology*. Amsterdam: John Benjamins, 79–105.
- Andersen, Roger W. and Yasuhiro Shirai. 1994. Discourse motivations for some cognitive acquisition principles. *Studies in Second Language Acquisition* 16/2: 133–156.
- Ayoun, Dalila and M. Rafael Salaberry. 2008. Acquisition of English tense-aspect morphology by advanced French instructed learners. *Language Learning* 58/3: 555–595.
- Bardovi-Harlig, Kathleen. 2000. *Tense and Aspect in Second Language Acquisition: Form, Meaning and Use*. Malden: Blackwell Publishers.
- Bardovi-Harlig, Kathleen and Anna Bergström. 1996. Acquisition of tense and aspect in second language and foreign language learning: Learner narratives in ESL and FFL. *Canadian Modern Language Review* 52/2: 308–330.

- Bardovi-Harlig, Kathleen and Llorenç Comajoan-Colomé. 2008. Order of acquisition and developmental readiness. In Bernard Spolsky and Francis M. Hult eds. *The Handbook of Educational Linguistics*. Malden: Blackwell Publishers, 383–397.
- Bardovi-Harlig, Kathleen and Llorenç Comajoan-Colomé. 2020. The aspect hypothesis and the acquisition of L2 past morphology in the last 20 years: A state-of-the-scholarship review. *Studies in Second Language Acquisition* 42/5: 1137–1167
- Bayley, R. 1994. Interlanguage variation and the quantitative paradigm: Past tense marking in Chinese-English. In Elaine Tarone, Susan Gass and Andrew Cohen eds. *Research Methodology in Second-Language Acquisition*. Hillsdale: Lawrence Erlbaum, 157–181.
- Bonilla, Carrie. L. 2013. Tense or aspect? A review of initial past tense marking and task conditions for beginning classroom learners of Spanish. *Hispania* 96/4: 624–639.
- Collins, Lauren, Joanna White, Pavel Trofimovich, Walcir Cardoso and Marlise Horst. 2012. When comprehensible input is not comprehensive input: A multi-dimensional analysis of instructional input in intensive English as a foreign language. In Carmen Muñoz ed. *Intensive Exposure Experiences in Second Language Learning*. Bristol: Multilingual Matters, 66–87.
- Comrie, Bernard. 1976. *Aspect: An Introduction to the Study of Verbal Aspect and Related Problems*. Cambridge: Cambridge University Press.
- Comrie, Bernard. 1985. *Tense*. Cambridge: Cambridge University Press.
- Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
- Daidone, Danielle. 2019. Preterite and imperfect in Spanish instructor oral input and Spanish language corpora. *Hispania* 102/1: 45–58.
- Davies, Mark. 2002. An annotated corpus of 100,000,000 words in historical and modern Spanish. *Sociedad Española para el Procesamiento del Lenguaje Natural* 29: 21–27.
- Domínguez, Laura, Nicole Tracy-Ventura, María J. Arche, Rosamond Mitchell and Florence Myles. 2013. The role of dynamic contrasts in the L2 acquisition of Spanish past tense morphology. *Bilingualism* 16/3: 558–577.
- Duchon, Andrew, Manuel Perea, Nuria Sebastián-Gallés, Antonia Martí and Manuel Carreiras. 2013. EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods* 45/4: 1246–1258.
- Ellis, Nick C. 2013. Frequency-based grammar and the acquisition of tense and aspect in L2 learning. In M. Rafael Salaberry and Llorenç Comajoan-Colomé eds. *Research Design and Methodology in Studies on L2 Tense and Aspect*. Berlin: Mouton de Gruyter, 89–118.
- Ellis, Nick C. and Fernando Ferreira-Junior. 2009. Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal* 93/3: 370–385.
- Fernández-Mira, Paloma, Emily Morgan, Samuel Davidson, Aaron Yamada, Agustina Carando, Kenji Sagae and Claudia Sánchez-Gutiérrez. 2021. Lexical diversity in an L2 Spanish learner corpus: The effect of topic-related variables. *International Journal of Learner Corpus Research* 7/2: 229–258.
- Goldberg, Adele. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7/5: 219–224.
- González, Paz and Lucía Quintana Hernández. 2018. Inherent aspect and L1 transfer in the L2 acquisition of Spanish grammatical aspect. *The Modern Language Journal* 102/3: 611–625.

- Gries, Stefan. 2014. *Coll.analysis 3.5. A Script for R to Compute Perform Collostructional Analyses*. <http://www.stgries.info/teaching/groningen/index.html>
- Gries, Stefan and Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9/1: 97–129.
- Izquierdo, Jesús and Maria Kihlstedt. 2019. L2 imperfective functions with verb types in written narratives: A cross-sectional study with Hispanophone learners of French. *The Modern Language Journal* 103/1: 291–307.
- Larsson, Tove, Jesse Egbert and Douglas Biber. 2022. On the status of statistical reporting versus linguistic description in corpus linguistics: A ten-year perspective. *Corpora* 17/1: 137–157.
- Lozano, Cristóbal. 2009. CEDEL2: Corpus escrito del español L2. In Carmen María Bretones Callejas, José Francisco Fernández Sánchez, José Ramón Ibáñez Ibáñez, María Elena García Sánchez, María Enriqueta Cortés de los Ríos, María Sagrario Salaberri Ramiro, María Soledad Cruz Martínez, Nobel Augusto Perdu Honeyman and Blasina Cantizano Márquez eds. *Applied Linguistics Now: Understanding Language and Mind*. Almería: Universidad de Almería, 197–212.
- Lozano, Cristóbal. 2021. CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. *Second Language Research*. <https://journals.sagepub.com/doi/full/10.1177/02676583211050522>
- Lozano, Cristóbal and Amaya Mendikoetxea. 2013. Learner corpora and SLA: The design and collection of CEDEL2. In Ana Díaz-Negrillo, Nicolas Ballier and Paul Thompson eds. *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam: John Benjamins, 65–100.
- MacWhinney, Brian. 2016. Entrenchment in second-language learning. In Hans-Jörg Schmidt ed. *Entrenchment and the Psychology of Language Learning*. Berlin: Mouton de Gruyter, 343–366.
- Padró, Lluís, Miquel Collado, Samuel Reese, Marina Lloberes and Irene Castellón. 2010. Freeling 2.1: Five years of open-source language processing tools. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner and Daniel Tapias eds. *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*. Valetta: European Language Resources Association, 931–936.
- Padró, Lluís and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the 8th International Conference on Language Resources and Evaluation Conference, LREC 2012*. Istanbul: European Language Resources Association, 2473–2479.
- Real Academia Española. *Corpus de Referencia del Español Actual (CREA)*. <http://corpus.rae.es/creanet.html>
- Robison, Richard. 1995. The aspect hypothesis revisited: A cross-sectional study of tense and aspect marking in interlanguage. *Applied Linguistics* 16/3: 344–370.
- Rosch, Eleanor and Carolyn B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7/4: 573–605.
- Salaberri, M. Rafael. 1999. The development of past tense verbal morphology in classroom L2 Spanish. *Applied Linguistics* 20/2: 151–178.
- Salaberri, M. Rafael. 2002. Tense and aspect in the selection of Spanish past tense verbal morphology. In M. Rafael Salaberri and Yasuhiro Shirai eds. *The L2 Acquisition of Tense-Aspect Morphology*. Amsterdam: John Benjamins, 397–416.

- Salaberry, M. Rafael. 2008. *Marking Past Tense in Second Language Acquisition: A Theoretical Model*. London: Continuum.
- Salaberry, M. Rafael. 2011. Assessing the effect of lexical aspect and grounding on the acquisition of L2 Spanish past tense morphology among L1 English speakers. *Bilingualism: Language and Cognition* 14/2: 184–202.
- Salaberry, M. Rafael and Dalila Ayoun. 2005. The development of L2 tense-aspect in the Romance languages. In M. Rafael Salaberry and Dalila Ayoun eds. *Tense and Aspect in Romance Languages: Theoretical and Applied Perspectives*. Amsterdam: John Benjamins, 1–33.
- Shirai, Yasuhiro. 2004. A multiple-factor account for the form-meaning connections in the acquisition of tense-aspect morphology. In Bill VanPatten, Jessica Williams, Susanne Rott and Mark Overstreet eds. *Form-Meaning Connections in Second Language Acquisition*. London: Routledge, 97–119.
- Thomas, Anita. 2014. Le rôle de l'aspect lexical et de la fréquence des formes dans l'input sur la production des formes du passé par des enfants apprenants du français L2 en début d'acquisition. *The Canadian Modern Language Review* 70/1: 1–27.
- Tracy-Ventura, Nicole. 2007. *Testing the Distributional Bias Hypothesis: A Corpus-based Study of Lexical and Grammatical Aspect in Spanish*. Southampton: University of Southampton dissertation.
- Tracy-Ventura, Nicole and John A. Cuesta Medina. 2018. Can native-speaker corpora help explain L2 acquisition of tense and aspect? A study of the 'input'. *International Journal of Learner Corpus Research* 4/2: 277–300.
- University of Wisconsin. 1998. *The University of Wisconsin College-Level Placement Test: Spanish (Grammar) Form 96M*. Madison: University of Wisconsin Press.
- Verkuyl, Henk. 1972. *On the Compositional Nature of the Aspects*. Dordrecht: Reidel.
- Wulff, Stephanie. 2020. Usage-based approaches. In Nicole Tracy-Ventura and Magali Paquot eds. *The Routledge Handbook of Second Language Acquisition and Corpora*. London: Routledge, 175–188.
- Wulff, Stephanie, Nick C. Ellis, Ute Römer, Kathleen Bardovi-Harlig and Chelsea J. Leblanc. 2009. The acquisition of tense-aspect: Converging evidence from corpora and telicity ratings. *The Modern Language Journal* 93/3: 354–369.
- Yamada, Aaron, Samuel Davidson, Paloma Fernández-Mira, Agustina Carando, Kenji Sagae and Claudia Sánchez-Gutiérrez. 2020. COWS-L2H: A corpus of Spanish learner writing. *Research in Corpus Linguistics* 8/1: 17–32.
- Zipf, George K. 1935. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Cambridge: MIT Press.

Corresponding author

Sophia Minnillo

469 Kerr Hall

University of California

One Shields Avenue

Davis, CA 95616

United States

E-mail: smminnillo@ucdavis.edu

received: January 2022

accepted: April 2022

Review of Pérez Paredes, Pascual. 2020. *Corpus Linguistics for Education: A Guide for Research*. London: Routledge. ISBN: 978-0-367-19843-5. <https://doi.org/10.4324/9780429243615>

Barry Pennock-Speck
University of València / Spain

This volume is the eighth in the series, *Routledge Corpus Linguistics Guide*, edited by Michael McCarthy and Anne O’Keeffe. In the description of the book, we are told that it “provides a practical and comprehensive introduction to the use of corpus research-methods in the field of education” and that the approach is “hands on.”¹ After a critical reading of the volume, I can state without hesitation that both assertions are true. *Corpus Linguistics for Education: A Guide for Research* is targeted at both students and researchers who are studying or carrying out research in Corpus Linguistics in the field of education. There are eight chapters including an introduction and a short conclusion. No previous knowledge of Corpus Linguistics is assumed as this volume in the series is explicitly described as an introductory textbook. The style is unpretentious and great efforts are made to make the volume readable. For example, Chapter 1 starts off in a light-hearted manner with “If you expect to find a definition of Corpus Linguistics in this opening paragraph, you will not be disappointed” (p. 1) and the author even makes a mouthful, such as “the typicality of hegemonic discourse” (Baker 2006), palatable by offering interesting quotes that demonstrate the usefulness of the term when related to disability and the press’s depiction of Islam and Muslims (p. 13). The 42 figures and 51 tables are designed to guide the reader through what is sometimes quite complex material. In Figure 1.4, for instance, the reader is shown how to differentiate a corpus as primary data from one used as secondary data. It is also used to further explain these two approaches and is then followed by further explanations. Tables are also often used

¹ <https://www.routledge.com/Corpus-Linguistics-for-Education-A-Guide-for-Research/Perez-Paredes/p/book/9780367198435>



in an explanatory fashion. For example, Table 1.3 shows the differences between positivism and phenomenology, the approach taken by researchers in each paradigm and the methods they use. Each chapter ends with a notes section made up of notes proper and the addresses of web sites, and a list of references. Both the links and the references are very useful for those who might want to look into aspects dealt with in each chapter in more depth. Competences are considered as well as knowledge through the 18 skills introduced chapter by chapter. In Chapter 4, skills one to 11 are reviewed, and skills 12 to 17 in Chapter 7. The book is complemented by a link to a document that includes suggested answers to the skill review questions.

The philosophy of Corpus Linguistics is summed up in the first paragraphs of Chapter 1 as the study of the language of real life and the empirical analysis of attested usage of actual language. The author then goes on to provide a definition of corpus, “a large body of texts” (p. 1), and further on describes a corpus as being both an instrument and a method designed to answer research questions. From an epistemological point of view, Corpus Linguistics is situated within the scientific paradigm, as it uses mainly quantitative methods and large samples, that is, corpora. There follows a description of several corpora and the research questions they helped to answer. The author outlines several important concepts in Corpus Linguistics such as accountability, falsifiability, replicability and representativeness using examples and quotes from leading authors in the field.

Chapter 2, the shortest of the chapters, focuses on text analysis in the field of education research. In the first part of the chapter, the author gives an account of the two main qualitative approaches employed in text analysis in the field, content analysis and theme analysis. He subsequently goes on to provide a brief description of software packages such as *NVivo*,² *MAXQDA*³ and *AntConc* (Anthony 2019) that can be used to code texts. We are then offered an overview of Conversation Analysis and Discourse Analysis —approaches that have used Corpus Linguistic tools to achieve their objectives (cf. Flowerdew 2012 or Walsh 2016). It is when we reach Section 2.2.1 that we are introduced to the first dissensions that exist in Corpus Linguistics, for example, corpus-based vs. corpus driven approaches and theory-driven and data-driven approaches. It is also after this section that we are introduced to quantitative and

² <http://image-analysis.com/>

³ <https://www.maxqda.com/>

qualitative approaches used alongside Corpus Linguistics. In the final section of the chapter, the focus is on the register perspective found in the work of Biber and Conrad (2009).

In Chapter 3, we are introduced to Corpus Linguistic approaches applied to the study of language use. Here the author underlines the importance of Corpus Linguistics to discover regularities and patterns in texts, which can help us to understand better the textual habits of communities. The author admits that, to a certain extent, Corpus Linguistics reduces the complexity of a text to a simpler form to make it more comprehensible. For the first time in the volume, we come across case studies. In the first study, we are introduced to both qualitative and Corpus Linguistic methods to analyse interviews, in the second to content analysis and Corpus Linguistics to examine educational policies. The case studies are useful in that, through practical examples, we are shown the differences and affordances of Corpus Linguistics when compared to qualitative approaches that employ interviews and thematic analysis mentioned in an earlier chapter. Through copious examples and explanations, in Section 3.2, we are given our first glimpse into the practicalities of Corpus Linguistics in the shape of concordance lines. The author includes a practical six-step procedure to analyse concordance lines. Technical terms, such as ‘lemmas’, ‘types’, ‘tokens’, ‘nodes’, are introduced at intervals, which, together with the screenshots from *AntConc*, *WordSmith* (Scott 2020) and *Sketch Engine* (Kilgarriff *et al.* 2014), makes it easier to grasp their meaning. The reader is also shown how the results from word lists can be exported. The handling of frequencies is explored next. Emphasis is placed on the importance of considering the size of corpora such as the *British National Corpus* (BNC; BNC Consortium 2007) or the *Corpus of Contemporary American English* (COCA; Davies 2008–) and the need to calculate a term’s relative frequency. As in the case of concordance lines and word lists, the reader is shown not only what a lemma is but how a lemma list is loaded into *AntConc* or where to go to download a lemma list (Mike Scott’s web site).⁴ Finally, definitions of *collocation* and *collocate* are provided and their importance for Corpus Linguistics is explained. Once more, figures and tables make it easier for the reader to understand the concepts that have been introduced.

⁴ <https://www.lexically.net/>

Chapter 4 focuses on designing corpora. The first section covers corpus size and data collection. The readers are warned against attempting to compile unrealistically large corpora. A corpus should be large enough to be representative of the type of language being studied and, importantly, while taking into account the time needed to compile it (Reppen 2010). The reader is then provided with two lengthy case studies. The first uses Corpus Linguistics as the main research methodology to analyse issues involving early childhood education in Australia. The reader is shown the research questions that were drawn up and the analyses performed to answer them. The second case study involves embedded Corpus Linguistic research methods. However, the reader is not given a definition of the meaning of ‘embedded Corpus Linguistics’. This study involves ‘narrative policy analysis’ and Corpus Linguistic methods to analyse aspects of literacy education in Canada after the crisis in 2008.

The second section, 4.2, goes into the basics of comparing corpora and significance testing but I have only found an indirect reference to significance testing in the reference to keyword analysis. The author puts forward that Corpus Linguistic research is frequently comparative and that, when two corpora are analysed, it is normal to gauge the differences in usage between two (or more) corpora or use a second one as a reference corpus to carry out a keyword analysis. To illustrate the comparison of corpora, the author looks at educational policy publications from the UK and New Zealand. Section 4.2.1 examines the functionality of Part-of-Speech (POS) tagging. This, the author states, adds sophistication to the searches as POS tagging can be combined with words or lemmas. The UK and New Zealand corpora are used to illustrate POS tagging. The numerous figures used to show POS tags in *Sketch Engine* and *AntConc* are extremely useful. Information is also given on freely available POS taggers that are freely available. The final sections of Chapter 4 are given over to a review of skills one to 11.

Chapter 5 is dedicated to describing the transcription and annotation of interview data. The author highlights the labour-intensive nature of transcribing interviews. Transcriptions, we are told, may just contain what was being said but could also include other details such as the tone of voice, inflection, emphasis, pauses, interruptions, etc. We are given a glimpse into the many decisions that need to be made when transcribing. The *Child Language Data Exchange System* (CHILDES)⁵ is offered as an example of

⁵ <https://childes.talkbank.org/>

thorough transcription and once more, as in earlier chapters, coding is brought up. In Section 5.2, basic transcription techniques are described. Two desktop solutions, *Inscribe*⁶ and *EXMARaLDA Partitur Editor*,⁷ designed to help researchers with transcriptions, are briefly outlined. The author adds that these can be complemented with software such as *Praat* (Hirst 2013), *ELAN* (Wittenburg *et al.* 2006) or the *UAM Corpus Tool*.⁸ Table 5.1 includes a comprehensive insight into the LINDSEI transcription guidelines.⁹ Readers are advised to employ setting brackets to tag annotations such as <foreign> and </foreign> to be able to find annotated text in software such as *AntConc*. Figure 5.2 is provided to illustrate how this is done. The final Section 5.3 emphasises the need to add structure and metadata to a corpus. 5.3.1 explains how to annotate a corpus with one's own tags to get the most out of searches using *Sketch Engine*. The final section, 5.3.2, deals with annotation using standard XML guidelines. The author suggests a *Text Encoding Initiative* (TEI) template to gather metadata (Pérez-Paredes and Alcaraz-Calero 2009) and goes into great detail in its description. Chapter 5 is not the longest chapter in the volume, that honour goes to Chapter 6, but to my mind it is the most complex. Nevertheless, great care is taken to make the explanations as clear and comprehensive as possible.

The remit of Chapter 6 is to provide the reader with insights into the Corpus Linguistic analysis of vocabulary. The keyword analysis of a corpus of peace treaties is employed to show how the concept of education is used in the texts. Keywords, as the author explains, can help to highlight the words that characterise the corpus under scrutiny. The reader is provided with a lengthy but very helpful step-by-step guide to keyword analysis using the corpus of peace treaties. In Section 6.3 there is a guide to researching nouns and noun phrases focusing on the examination of their colligational behaviour using *Sketch Engine*, first by focusing on individual nouns, in this case *education*, and then multiword units. The final section, 6.4, on the lexicon of children's literature involves various corpora. It is here that we come across N-Grams, which are described in depth for the first time. Strangely, N-Grams are not highlighted as essential terminology as are other terms, but Table 6.11 offers a summary of how they are used.

⁶ <https://www.inqscribe.com/>

⁷ <https://exmaralda.org/en/partitur-editor-en/>

⁸ <http://www.corpustool.com/index.html>

⁹ <https://uclouvain.be/en/research-institutes/ilc/cecl/transcription-guidelines.html>

Chapter 7 centres on examining talk and how to tease out the collocations and patterns in conversations and interviews. Corpus Linguistics, the author states, gives researchers unique insights into how language is used. Using the *Backbone Corpus of English as a Lingua Franca* (Kohn 2012), readers are shown how Corpus Linguistic methods can serve to carry out more complex searches. There follows a review of the major differences between spoken and written language. The systemic functional grammar approach is used (Locke 2004), for example, to classify transitivity. Section 7.2 outlines how to do multiword keyword searches through Corpus Query Language (CQL) in *Sketch Engine*. In Section 7.2.3 readers are instructed on how to run a search that will show how family life is impacted by work in the *Backbone Corpus*. This is possible as the coders had included the information in the corpus. The author links the findings to the principles talked about in Chapter 1. Section 7.3 reviews skills 12 to 17, that is, those found in Chapters 5 to 7.

Chapter 8, the conclusion, is short but emphasizes the positivist nature of Corpus Linguistics and finishes with skill 18, that is, remaining critical. Here the reader is warned to be aware that what they do with Corpus Linguistics depends on the design of the corpus, the methods used and the interpretation of the results.

All in all, *Corpus Linguistics for Education: A Guide for Research* embodies a model introductory textbook. Very difficult concepts are introduced and fleshed out in a logical and straightforward manner. More importantly, the theory is linked, at all times, to the actual practice of Corpus Linguistics. This process is helped along greatly by the generous offerings of figures, tables, links to websites and copious references to seminal publications. As the readers are shown the workings of some of the most popular Corpus Linguistic programmes, and are given the links to several freely downloadable corpora, I imagine that several would be sorely tempted to try their hand at Corpus Linguistic analysis even before finishing the book—as I did.

REFERENCES

- Anthony, Laurence. 2019. *Antconc* (version 3.5.8). Tokyo, Japan: Waseda University.
<https://www.laurenceanthony.net/software/antconc/>
- Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.
- Biber, Douglas and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

- BNC Consortium. 2007. *The British National Corpus*, version 3 (BNC XML Edition). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. <http://www.natcorp.ox.ac.uk/>
- Davies, Mark. 2008–. *Corpus of Contemporary American English*. <https://www.english-corpora.org>
- Flowerdew, Lynne. 2012. Corpus-based discourse analysis. In James Gee and Michael Handford eds. *The Routledge Handbook of Discourse Analysis*. London: Routledge, 174–187.
- Hirst, Daniel. 2013. Anonymising long sounds for prosodic research. In Brigitte Bigi and Daniel Hirst eds. *Tools and Resources for the Analysis of Speech Prosody*. Aix-en-Provence: Laboratoire Parole et Langage, 36–37.
- Kilgariff, 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.
- Kohn, Kurt. 2012. Pedagogic corpora for content and language integrated learning. Insights from the BACKBONE Project. *The Eurocall Review* 20/2: 3–22.
- Locke, Terry. 2004. *Critical Discourse Analysis*. London: Bloomsbury.
- Pérez-Paredes, Pascual and José M. Alcaraz-Calero. 2009. Developing annotation solutions for online data driven learning. *ReCALL* 21/1: 55–75.
- Reppen, Randi. 2010. Building a corpus: What are the key considerations? In Anne O’Keeffe and Michael McCarthy eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 31–37.
- Scott, Michael. 2020. *WordSmith Tools version 8*. Stroud: Lexical Analysis Software.
- Walsh, Steve. 2016. Applying corpus linguistics and conversation analysis in the investigation of small group teaching in higher education. In Halina Chodkiewicz, Piotr Steinbrich and Malgorzata Krzeminska-Adamek eds. *Working with Text and Around Text in Foreign Language Environments*. Bern: Springer, 205–222.
- Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk and Daniel Tapias eds. *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*, 1556–1559.

Reviewed by
 Barry Pennock-Speck
 University of València
 Faculty of Philology, Translation and Communication
 Department of English and German
 46010. Valencia
 Spain
 E-mail: barry.pennock@uv.es

Review of Bouso, Tamara. 2021. *Changes in Argument Structure: The Transitivity Reaction Object Construction*. Bern: Peter Lang. ISBN: 978-3-034-34095-3. <https://doi.org/10.3726/b17960>

Sune Gregersen
University of Copenhagen / Denmark

1. INTRODUCTION

Tamara Bouso's *Changes in Argument Structure* is a book-length study of an intriguing phenomenon in English, the so-called Reaction Object Construction (ROC). An example of this construction is given in (1), the first of many examples discussed in this book:

(1) She mumbled her adoration.

In this and other examples of the ROC, a prototypically intransitive verb (in this case *mumble*) is used with a direct object (*her adoration*) expressing “a reaction or an attitude of some kind” (p. 15). In her book, Bouso sets out to provide a comprehensive account of this construction from the perspective of Construction Grammar, and to investigate its historical development with a particular focus on the Late Modern English period.

The book consists of eight chapters. After the introductory Chapter 1, these are grouped into two main parts. Part I, “Transitivity, Reaction Objects, and Construction Grammar” (Chapters 2–4), provides a comprehensive treatment of earlier literature on the subject and lays the theoretical foundation for Bouso's investigation. Part II is titled “Hands-on with data: A usage-based approach to the history of the ROC” (Chapters 5–8) and presents the empirical study of the ROC, focusing on its development and its relation to other constructions.



2. SUMMARY

Chapter 1 provides a first overview of the ROC and its place among other valency-changing constructions in English. Bouso defines the ROC as a type of argument augmentation (or valency-increasing) construction where an extra argument is added to the argument structure of the verb; other examples of such constructions include the cognate object construction (*She smiled an enigmatic smile*) and the *way*-construction (*She worked her way to the top*). After this overview, the main goals and research questions of the book are presented. These concern the analysis of the construction from a Construction Grammar perspective and the timing and causes of its development. Bouso formulates two diachronic hypotheses which are to be tested by the empirical study. The first is that the “formal restriction of coreferentiality of ROCs” (p. 27) goes back to the Early Modern English period. (I return to Bouso’s term ‘coreferentiality’ below). The second is that the ROC follows a diachronic trajectory similar to the cognate object and *way*-constructions, “occurring first with more transitive-like verbs and then expanding to intransitives” (p. 27).

In Chapter 2, “The process of transitivity in the history of English,” Bouso presents a brief survey of earlier scholarship on valency increase in the history of English. The point of departure is Visser (1963–73), whose observations on the matter are compared to those of some more recent scholars. Bouso points to a general consensus in the literature that English has developed an increasing number of ‘amphibious’ verbs (Visser’s term), that is, verbs which may be used both transitively and intransitively. In contrast, Old English had a larger number of verbs which were exclusively used intransitively (as far as we can tell from the sources, one might add). Some of the possible factors leading to this general development are then discussed, including morphological losses, ambiguity between *be*-perfects and *be*-passives in earlier English, and the creative use of novel reported speech verbs among some writers beginning in Late Modern English.

Chapter 3 is titled “Reaction objects: Review of the literature.” It begins with an overview of some earlier literature on types of objects in English —particular attention is paid to Jespersen (1909–49)— before moving on to the ROC and the two allegedly related constructions mentioned above, the cognate object construction and the *way*-construction. The similarities and differences between these constructions are examined at some length, though Bouso takes the position that the ROC is actually more closely

related to another phenomenon, effected objects, that is, objects whose referents come into being through the verbal activity, as in (2) and (3):

(2) She wrote a story.

(3) They dug a grave.

The similarities between effected objects and reaction objects are said to include their non-occurrence in resultative constructions (**They dug a grave rough*, **He smiled his welcome noticeable*), the impossibility of converting them to middle subjects (**The story wrote easily*, **The adoration mumbled easily*, etc.), and their non-occurrence with the definite article (I will return to this point below). The states of affairs expressed also tend to be inherently telic in both constructions, so an example like *She sang her thanks in an hour* is judged to be correct, while **She sang her thanks for an hour* is not (p. 94).

In Chapter 4, we get an overview of “Construction Grammar: Synchronic and diachronic perspectives.” First the central tenets of Construction Grammar are introduced with brief discussions of some well-known examples from the literature, such as *let alone* and the Preposition + Noun construction (i.e. *at work*, *in prison*, etc.) discussed by Goldberg (2013). After this some recent diachronic works from a Construction Grammar perspective are discussed. Some of the most important notions here include the distinction between ‘constructionalization’ and ‘constructional change’ proposed by Traugott and Trousdale (2013), the concept of a constructional network linking the constructions of a language to each other, and the idea that a given construction may ‘inherit’ features from several more schematic or abstract constructions (multiple inheritance).

Chapter 5, “The formation of ROCs,” begins the empirical part of the book. This chapter consists of two more or less independent sections: Section 5.1, “Characterization of the ROC,” and Section 5.2, “On the emergence of the ROC.” Section 5.1 concerns the analysis of the Present-day English ROC in Construction Grammar terms. Bouso first argues, I think convincingly, that the ROC should be treated as a construction in its own right, and then discusses its grammatical characteristics and a number of subtypes. A three-way typology proposed by Martínez-Vázquez (2015) appears to be particularly useful. Martínez-Vázquez distinguishes between ROCs with ‘delocutive’, ‘deverbal illocutionary’, and ‘predicative expressive’ objects. I illustrate these with three of Bouso’s examples in (4)–(6), respectively:

- (4) She waved him an adieu. (Thackery, *Vanity Fair*, cited p. 132) – DELOCUTIVE
- (5) The Chief Justice smiled acquiescence. (Darwin, *The Voyage of the Beagle*, cited p. 134) – DEVERBAL ILLOCUTIONARY
- (6) Mistress Grofe sat at her end of the table and glared her anger at all of us. (Taken from COCA, cited p. 136) – PREDICATIVE EXPRESSIVE

In the first of these, the reaction object refers to a greeting or other conversational routine, such as *an adieu* in (4). In the second, the reaction object is derived from a speech-act verb and may often be paraphrased as a verbal expression instead (*The Chief Justice acquiesced by smiling*). In the third type, the reaction object refers to the mental state of the subject and often has “adjectival features” (p. 136), though it is not necessarily deadjectival (compare Bouso’s examples *delight* and *joy*). In addition to this three-way typology, which makes reference to the type of reaction object, Bouso argues for a distinction between ROCs with and without an overt or implied recipient. ROCs without a recipient are monotransitive, whereas ROCs with a recipient have an additional participant and are similar in structure to a prototypical ditransitive construction. This is most obvious in cases like (4), but the recipient may also be implied in the context; according to Bouso, (5) is an example of this (see pp. 134, 141). Section 5.1 ends with a list of the most important grammatical properties of the ROC and a discussion of its relation to other constructions in the Present-day English ‘constructional network’: ditransitives, resultatives, and the monotransitive experiencer construction. Bouso argues that the ROC is a hybrid construction which inherits features from all of these.

Section 5.2 then traces the origin of the ROC in historical sources. Taking the works of Visser (1963–73), Jespersen (1909–49), and Levin (1993) as her point of departure, Bouso compiles an overview of all verbs mentioned in these sources which occur in the ROC. In very comprehensive tables (often running across several pages), she gives information on the earliest attestation and examples of all these verbs. In total, 69 verbs occurring in the ROC were identified in Visser, Jespersen, and Levin, to which 12 verbs were added from the *Oxford English Dictionary* (OED). A few of these verbs are attested in the ROC already in Middle English (namely *moan*, *bray*, *yelp*, and *roar*), but the vast majority are first attested in the construction in Late Modern English. Bouso compares the emergence of the ROC to the development of the cognate object

construction, the *way*-construction, and (very briefly) the ‘dummy *it*’ object construction (*snake legs it to freedom*; Mondorf 2016). She argues that these constructions have all followed a similar trajectory, expanding to an increasing number of intransitive verbs in Modern English.

After this follow two chapters devoted to the “Development of the ROC in British English” (Chapter 6) and the “Development of the ROC in American English” (Chapter 7). The two chapters not only investigate the ROC in two different written varieties, but also tackle rather different questions about the construction. Chapter 6 focuses on British English in the Late Modern English period (1710–1920) and takes as its point of departure 40 of the verbs which, in Chapter 5, were found to occur in the ROC. Bouso investigates which of these are attested in the ROC in *The Corpus of Late Modern English Texts* (CLMET3.0; De Smet *et al.* 2011), which reaction objects they occur with, and how strongly they are associated with the construction. A collexeme analysis reveals that a number of verbs are particularly often found in the construction, including *mutter*, *murmur*, and *smile*. An overview is also provided of all examples of non-human (i.e. animal or inanimate) subjects found among Bouso’s corpus results, and a number of individual verbs are discussed at greater length, such as *smile* and *nod*. Finally, Bouso considers the role of text type in the development of the construction. She notes that it is particularly frequent in narrative texts and that its peak in frequency in the middle of the investigated period (c. 1780–1850) coincides with the flourishing of the sentimental novel, where it seems to have been a favoured stylistic device.

Chapter 7 uses the *Corpus of Historical American English* (COHA; Davis 2010–), to investigate the development of the ROC in American English in the period 1810–2009, with particular attention to the productivity of the construction. Here a number of common reaction objects serve as the starting point rather than the verbs found in the construction. Bouso searches the corpus for a number of delocutive nouns which often occur in the construction (*hello(s)*, *goodbye(s)*, *thank you*, and a few variants) and manually identifies all instances of the ROC. 80 different verbs are attested with these reaction objects, which Bouso groups into six types: sound emission (*bark*), gesture (*nod*), bodily processes (*snuffle*), instrument of communication (*phone*), activity (*dance*), and light emission (*flare*). The findings suggest that the ROC has become increasingly productive in American English throughout the period, only peaking in the second half of the twentieth century. Bouso cautiously suggests that the ROC may

initially have been a predominantly British phenomenon, which was only imported into American English in the nineteenth century. The changing productivity of the ROC is then compared to observations made in earlier investigations of the more well-known *way*-construction.

Finally, Chapter 8 offers a “Summary and conclusion.” The summary first reiterates the main points of the individual chapters and then presents a brief sketch of the history of the ROC, distilling the many empirical observations into a condensed narrative. Bouso concludes that the ROC became a construction in its own right in Early Modern English after a number of ‘pre-constructionalization’ developments. After its constructionalization, the ROC increased its frequency considerably in Late Modern English. Bouso attributes this to language-internal factors, such as an alleged general increase in transitive verbs, and to language-external factors like the aforementioned development of the sentimental novel. The chapter ends with some suggestions for further research. The backmatter of the book contains lists of tables, figures, and references, as well as a short abstract.

3. DISCUSSION

This is a thought-provoking and empirically rich study, which sheds new light on an overlooked construction and its history. One gets the sense that few stones have been left unturned in Bouso’s work on the ROC. The review of the existing literature on the construction is very comprehensive, and the book contains more than 450 numbered examples in total, so it may be used both as a bibliography of earlier work and a handy data source for other scholars interested in the ROC. In addition, the empirical part of the book (Chapters 5–7) showcases how the same construction may be explored from various angles and with different methods, most of which may readily be applied to other corpora (e.g. covering other historical periods or other languages/varieties). While a number of remarks in the following will be of a more critical nature, in particular concerning some of Bouso’s analytical choices, it should be clear from the outset that the book is a useful addition to the literature in several respects.

This being said, I think there are a number of problems with Bouso’s characterization of the ROC and its place within the grammar of English. It is worth discussing this point at some length, as one of the stated goals of the book (pp. 25–26) is

to characterize the construction and its place within the English constructional network. While some of the potential issues may mainly be due to unclear or nonstandard terminology, I believe others are more fundamental. In some respects, in fact, I think Bouso's account of the ROC is contradicted by her corpus data. I begin the discussion by quoting Bouso's schematic representation of the ROC in (7) and her prose description of the construction in (8), both from p. 146:

(7) Syntax: SUBJ_i [V_{INTR}manner/means (OBJ1) OBJ2_i]. Where OBJ2 = (POSS)_i NP
Semantics: 'Sentient agent_i cause Y_i become expressed while/by_{manner/means} doing V'.

(8) the subject is an experiencer or sentient agent, as derived from the expressive meaning of the ROC as a whole; OBJ2, the reaction object proper, is an object of result which is coreferential with the subject; OBJ1 represents the recipient, which does not need to be always explicit, hence the notation in parentheses. Finally, V is an intransitive verb coding means or manner.

This characterization of the construction is referred to at several points in the book, and the formalization in (7) is repeated at least three times (see pp. 272, 316, 321). However, I believe a number of aspects of the description call for critical remarks. Firstly, note the slippery use of the referentiality index 'i' and the term 'coreferential'. In (7), the index is first added to the reaction object as a whole (OBJ2_i), then only to the optional possessive pronoun in the object NP [(POSS)_i]. In the description in (8), it is explicitly stated that the reaction object "is coreferential with the subject." This use of the term 'coreferential(ity)', which is repeated several times throughout the book (e.g. pp. 144, 159, 253–254), is highly unorthodox. In its received sense (see e.g. Trask 1993: 64–65 or Crystal 2008: 116–117), this term is used to refer to linguistic expressions which have the same extralinguistic referent, such as the pronouns *she* and *her* in (9):

(9) She_i mumbled her_i adoration.

In such examples, Bouso uses the referentiality index in accordance with the tradition. However, the book also contains numerous examples like (10):

(10) Pigs_i squeal emphatic disapproval_i. (p. 27)

Here the subject and the reaction object as a whole are said to be coreferential, in line with Bouso's description in (8). But there is no coreferentiality here in any received sense of the term, as the NPs *Pigs* and *emphatic disapproval* do not have the same extralinguistic referent. What Bouso seems to mean is that there is a close connection

between the subject and the reaction object, but the nature of this connection is never explored, and it is not discussed how (or even acknowledged that) Bouso's use of 'coreferential(ity)' differs from the linguistic tradition. This is more than just a minor terminological issue, since coreferentiality is taken to be a defining feature of the ROC. In addition, this terminological inaccuracy means that an opportunity is missed to provide a more exact characterization of the relation between the subject and the reaction object. To me, it would seem to be much better described as a type of possession (in a rather broad sense), as suggested by the frequent appearance of a possessive pronoun in the object NP, though this description may not apply equally well in all instances. In the delocutive type in particular, another label might be more appropriate.¹

Secondly, it is worth noting that the description of the subject as "an experiencer or sentient agent" is not entirely accurate, as Bouso herself points out later in the book. Inanimate subjects do in fact occur in the material (p. 231), although Bouso is of course correct that many of these are instances of metonymy, metaphor, or personification (as in *the Earth has just whispered a warning* from Shelley). This is not always obvious, however. In the example *The door jingled a welcome*, which is repeated at several points, I fail to see how the verb is used "metaphorically" (p. 130). Hence, even if clear examples of inanimate subjects may be rare in the ROC, it seems to me somewhat beside the mark to include animacy as a defining feature and formulate a separate "animacy constraint" on the construction, as Bouso does on p. 144.

Thirdly, a rather surprising aspect of Bouso's account is her characterization of the reaction object as an "object of result" (as in (8) above) or an "*effectum* object" (p. 92), that is, an object whose referent comes into being because of the verbal activity. This is surprising because so many of Bouso's own examples seem at odds with the description, in particular those belonging to the subtype of 'predicative expressive' ROCs, such as (11)–(12):

(11) She smiled disbelief. (p. 165)

(12) He only grunted his gratitude. (p. 166)

¹ For instance, in examples like *The door jingled a welcome* (p. 15 and elsewhere) or *The girls wave a farewell to the men* (p. 259), where it is not obvious that the reaction objects *a welcome* and *a farewell* are really possessed by the subject referents. Still, 'possession' would be a more apt term than 'coreferentiality' even in these cases.

In these examples, the disbelief and gratitude are surely not created by the verbal activities, but rather communicated (or “expressed,” as Bouso puts it in (7)). Similar considerations apply to many other examples given in the book, such as the ones quoted in (6), (9), and (10) above. The analysis of reaction objects as objects of result appears to be due mainly to Martínez-Vázquez (1998) and Kogusuri (2009), but I think the similarities between these two types of objects are overstated both by these authors and by Bouso (e.g. on pp. 93–95). It is certainly not the case that objects of result cannot contain a definite article, as Bouso claims. One of her own examples contains a definite article (*The dressmaker made the dress*, p. 58), and one may easily find additional examples like (13), from the OED (s.v. *write* v. 14b):

- (13) But the poems are harmless. Love poems. And diaries. You wrote the poems for your girls, isn't it? (2002 H. Habila *Waiting for Angel* (2003) 16)

In fact, the two starred examples given by Bouso to show that objects of result are incompatible with the definite article (**I dig the grave*, **She lights the fire*, p. 95) return numerous hits on Google and are attested verbatim in the COHA:

- (14) There I will bury him, if I dig the grave myself. (COHA, 1918 FIC)

- (15) She lights the fire and puts more coffee in the pot. (COHA, 1914 FIC)

A puzzle thus remains about the ROC, which indeed appears never to contain a definite article in the object NP. Even if some (but not all) reaction objects may be analyzed as objects of result, this cannot explain the constraint, as there is no general restriction on definite articles in objects of result. Here an open question thus remains for future research.²

Finally, I note that Bouso's initial definition and description in (7)–(8) does not explicitly mention particle verbs, although examples with the particles *forth*, *off*, *out*, *over*, and *up* are later included in the analysis of the British English material. Indeed, a revised version of (7) which explicitly mentions these particles is provided later (p. 213). This is an important point because such particles are said to have played a role in

² Another open question is whether the ROC is regularly used in negative contexts in Present-day English. At one point, Bouso mentions in passing that reaction objects “can never occur in negative or interrogative sentences” (p. 92). But note examples like (i)–(ii), both retrieved from Google Books:

- (i) No bell jingled a welcome as she stepped into the dimly lit interior and peered around. (Cox, *Trouble in Store*, 2013)
- (ii) He didn't nod his thanks but gave her a thumbs up. (Campbell, *Catawba Point*, 2020)

the development of the construction. A number of examples like (16) are cited in the discussion of the history of the ROC:

- (16) With eies which glistered forth beames of disdaine. (Sir P. Sidney, *Arcadia* (1590), cited on p. 157)

Bouso finds that “14 of the 51 earliest attested instances of ROCs from Levin’s list feature the particles *forth*, *out*, and *up*, which in some way reassures the transitivizing power of these particles” (pp. 203–204). The exact nature of this ‘transitivizing power’ is not explored further, however, and it is unclear what role particle verbs play in the ROC in Present-day English.

One very attractive aspect of the book is that it paves the way for future studies in numerous respects. Bouso repeatedly acknowledges that her book is not the final word on the topic (see e.g. pp. 16, 147, 276) and makes several interesting suggestions for future research. Among many other topics, these include the role of genre in the development of the construction, where Bouso suggests that the popularity of the sentimental novel may have been especially important. A crucial question here is whether the ROC is primarily or even exclusively a literary phenomenon, or whether the construction was always productive in the colloquial language, but merely happens to have a higher frequency in specific literary genres. Bouso clearly prefers the former interpretation and even suggests that “to judge from the low frequency of occurrence of ROCs and their restriction to specific text types [...] most likely ROCs are not part of the construction or linguistic knowledge of a large set of [the] population” (p. 265).³ This is an interesting suggestion, but a more targeted investigation of more colloquial genres, both historical and contemporary, would be necessary to substantiate (or challenge) it.

Another topic which deserves more attention is the earliest history of the construction, as Bouso herself states in the conclusion (pp. 320–321). As mentioned above, the empirical part of the book mainly focusses on Late Modern English, which may seem somewhat surprising given Bouso’s conclusion that the ROC was constructionalized already in the Early Modern English period. I thus agree with Bouso

³ Presumably, “linguistic knowledge” in this quotation should be understood as *active* linguistic knowledge. Given the use of the ROC in many literary texts, one must assume that most if not all native speakers at least have *passive* knowledge of the construction. For future investigations of the creative use of the construction, I note in passing that the ‘syntactic blend’ approach of Hampe and Schönefeld (2003, 2007), which is not referred to by Bouso, might provide an interesting perspective.

that a more detailed investigation of the medieval and Early Modern English situation is necessary to get a fuller understanding of how the construction emerged. In addition, I think the criteria for constructionalization need to be made more explicit if this notion is to be of much value. Bouso argues that the ROC was constructionalized —i.e. became “a *new* form-meaning pairing” (p. 26)— in Early Modern English. The main arguments for this appear to be that the constraint of ‘coreferentiality’ (or rather ‘possession’, as I argued above) developed in Early Modern English, and that the construction became productive with an increasing number of intransitive verbs in this period (see pp. 317, 320). Bouso indeed finds only four Middle English examples of the ROC in the sources (all of them in verse texts), whereas there are numerous attestations in the Early Modern English material. However, without a more principled quantitative investigation, it is unclear just how much the productivity increased in Early Modern English. Note also that the four alleged Middle English examples all have a possessed object, as shown in (17)–(20):

- (17) And don h[i]m monen his sinfulhed
 ‘and make him bemoan his sinfulness’ (c1250 *Genesis and Exodus* l. 180)
- (18) Braundysch & bray by brabez breme
 ‘[though you] struggle and cry out your violent rage’ (c1400 *Pearl* l. 346)
- (19) His sorwe coude he to no man zelpe
 ‘His sorrow he could call out to no one’ (c1400 *Laud Troy Book* 13520)
- (20) Mi bollid hert doth so his sikis rore /
 that mawgre me hit doth my wele biwray
 ‘My swollen heart cries out its sighs so /
 that in spite of myself it betrays my will’
 (c1450 *Charles d’Orleans Poems* 219)⁴

In other words, the Middle English examples identified by Bouso all appear to satisfy her constraint of ‘coreferentiality’. It is thus not clear to me why this constraint is only said to have developed in the Early Modern English period.⁵ On this point, future work will hopefully provide some clarification as well.

⁴ Note that (20) might also be analyzed as a causative (i.e. ‘My swollen heart so makes its sighs cry out’), but that the expression *sikis rore* occurs elsewhere in the same text (cf. Steele 1941: 191). It is not clear to me, though, why either (17) or (20) count as ROCs. In (17) *monen* means ‘bemoan, regret’, not ‘communicate by moaning’; *his sikis* ‘its sighs’ (see OED, s.v. *sike* n.²) in (20) is at least marginal and does not seem to fit comfortably in any of Bouso’s subcategories (see (4)–(6) above).

⁵ Another question is why the feature of ‘coreferentiality’ in particular should be taken as a diagnostic of constructionalization rather than, say, the semantics of the object noun. On the problem of diagnosing

Stylistic infelicities and typographical errors are few and far between in the book, and the writing is generally transparent and easy to follow. In several places, however, the presentation of the material could have been more reader-friendly. Numbered examples are occasionally presented *en bloc* rather than one by one when they are discussed in the text, meaning that one has to go back and forth between the examples and Bouso's discussion of them (see e.g. pp. 87–89, 240–242, 297–302). On a related note, the linguistic data are sometimes presented in rather unwieldy tables running across several pages (e.g. pp. 160–169, 171–185). These are minor nuisances, however, in an interesting study touching on some crucial issues in English grammar. Although I have questioned some of Bouso's analytical choices in the above discussion, it should be clear enough that her book is essential reading for anyone interested in the English reaction object construction. It will hopefully inspire more work on this particular construction in Middle and Early Modern English or from a stylistic or sociolinguistic perspective, and on related phenomena in other languages of the world.

REFERENCES

- Börjars, Kersti, Nigel Vincent and George Walkden. 2015. On constructing a theory of grammatical change. *Transactions of the Philological Society* 113/3: 363–382.
- Crystal, David. 2008. *A Dictionary of Linguistics and Phonetics*. Malden: Blackwell.
- Davies, Mark. 2010–. *Corpus of Historical American English*. <https://www.english-corpora.org>
- Goldberg, Adele E. 2013. Constructionist approaches. In Thomas Hoffmann and Graeme Trousdale eds. *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press, 15–31.
- Gregersen, Sune. 2018. Some (critical) questions for diachronic construction grammar. *Folia Linguistica Historica* 39/2: 341–360.
- Hampe, Beate and Doris Schönefeld. 2003. Creative syntax: Iconic principles within the symbolic. In Wolfgang G. Müller and Olga Fischer eds. *From Sign to Signing*. Amsterdam: John Benjamins, 243–226.
- Hampe, Beate and Doris Schönefeld. 2007. Syntactic leaps or lexical variation? – More on “Creative Syntax.” In Stefan Th. Gries and Anatol Stefanowitsch eds. *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*. Berlin: Mouton de Gruyter, 127–157.
- Hendrik De Smet, Hans-Jürgen Diller and Jukka Tyrkkö. 2011. *The Corpus of Late Modern English Texts*, version 3.0. <https://perswww.kuleuven.be/~u0044428/>
- Jespersen, Otto. 1909–49. *A Modern English Grammar on Historical Principles*. Copenhagen: Munksgaard.

- Kogusuri, Tetsuya. 2009. The syntax and semantics of reaction object constructions in English. *Tsukuba English Studies* 28: 33–53.
- Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.
- Martínez-Vázquez, Montserrat. 1998. Effected objects in English and Spanish. *Languages in Contrast* 1/2: 245–264.
- Martínez-Vázquez, Montserrat. 2015. Nominalized expressive acts in English. *Verbum* 37/1: 147–170.
- Mondorf, Britta. 2016. “Snake legs it to freedom”: Dummy *it* as pseudo-object. *Corpus Linguistics and Linguistic Theory* 12/1: 73–102.
- Oxford English Dictionary Online*. Oxford: Oxford University Press.
<https://www.oed.com/>
- Steele, Robert. 1941. *The English Poems of Charles of Orleans*. London: Oxford University Press.
- Trask, R. Larry. 1993. *A Dictionary of Grammatical Terms in Linguistics*. London: Routledge.
- Traugott, Elizabeth Closs and Graeme Trousdale. 2013. *Constructionalization and Constructional Changes*. Oxford: Oxford University Press.
- Visser, Fredericus Th. 1963–73. *An Historical Syntax of the English Language*. Leiden: Brill.

Reviewed by
 Sune Gregersen
 University of Copenhagen
 Department of Nordic Studies and Linguistics
 Emil Holms Kanal 2
 2300 København S
 Denmark
 E-mail: s.gregersen@hum.ku.dk

Review of Moskovich, Isabel, Inés Lareo and Gonzalo Camiña. 2021. *“All Families and Genera”: Exploring the Corpus of English Life Sciences Texts*. Amsterdam: John Benjamins. ISBN: 978-9-027-20924-5. <https://doi.org/10.1075/z.237>

Stefania Degaetano-Ortlieb
Saarland University / Germany

Historical linguistics is witnessing a major shift since the early twenty-first century towards the integration of quantitative approaches in the methodological repertoire of the discipline. As Jensen and McGillivray (2017) observe, while the shift towards quantitative methods has penetrated many subfields of linguistics already, historical linguistics has only recently boosted interdisciplinary collaboration with quantitative linguistics. Indispensable for this endeavor is corpus-based work, which has been taken very seriously in this book by Isabel Moskowich, Inés Lareo and Gonzalo Camiña in building the *Corpus of English Life Sciences Texts*¹ (CELiST). From the collection of papers of this book, the interdisciplinary work becomes evident, including not only transparency in corpus-building decisions accounting for the history of science, but also evaluation procedures on the corpus and its representativeness in light of the integration of sound knowledge on the history of science, as well as insights gained from working with the corpus. The chapters of the book fall roughly in the above mentioned three parts. In Chapters 1 to 4, the making of the corpus, the editorial policy adopted to select and encode material and a detailed description of the eighteenth and nineteenth century samples are presented. Chapter 5 evaluates in a detailed manner the representativeness of the corpus, whereas Chapters 6 to 15 present corpus-based studies on various linguistic aspects of the CELiST texts ranging from lexical variation to discourse

¹ <https://varieng.helsinki.fi/CoRD/corpora/CELiST/>



matters up to register-internal shifts, most of the analytical contributions focusing on evaluative language.

Considering the first part of the book, the first chapter introduces the CELiST corpus in terms of the fields of natural science covered and how the authors decided to group these under the broader term ‘Life Sciences’. The authors are very transparent about their selection procedure (something that is faithfully continued throughout the first part of the book). The corpus amounts at 10,000-word samples per decade with a total of 400,305 words. Given the endeavors put into corpus quality, this is quite an achievement. Moreover, the corpus is rich in metadata reflecting the socio-historical context of both time periods covered (eighteenth and nineteenth centuries). The metadata encompasses authors, authors’ gender, genres, as well as geographical information, that is, information about the English-speaking country in which the authors were educated and acquired their linguistic habit. For the latter, on page 12 there is a mismatch between the description of increase of Ireland-related authors when comparing Figure 4 to 5: in fact, it decreases, while percentage of authors educated in North America increases from the eighteenth to the nineteenth centuries. Also missing in this chapter is the description of authors’ age, a variable used in some of the analyses. The second chapter describes, in detail, the editorial decisions taken by the book authors to encode the corpus material. The challenges described, which are well-known by historical linguists working with digital material, are remarkably dealt with. The effort was not limited to one aspect but went into OCR-error correction up to visualizing the texts as authentically as possible, and particularly highlighted should be the endeavor to truthfully reflect the authors language giving the possibility to exclude language of others present in the texts, such as quotes. Chapters 3 and 4 are summarizations of the content of the texts represented in the corpus, for the eighteenth and nineteenth centuries respectively, however with details on the texts’ length and document structure and most importantly on the socio-historical context, which is especially relevant to the historical linguist. Collecting this kind of information would mean quite some work and having that covered in the book is of great value.

In the second part, Chapter 5 presents what is an often-missed contribution in books about corpora, namely corpus evaluation. It provides a detailed computational evaluation of the representativeness of CELiST, confirming the corpus adequacy to represent the Late Modern English scientific discourse in a satisfactory way.

Analytically, the chapter presents results on how CELiST shows a constant lexical growth in line with specialization processes shaping scientific English at that time.

The third part of the book, dealing with analyses of the corpus, is introduced by looking at the lexical fixedness within CELiST considering binomials (such as *more or less*), their distributions across time and genres, as well as semantic relations between the components of the binomials. Nominal pairs are by far the most frequent binomials and unsurprisingly especially those joined by *and*. Diachronically, the distribution of binomials seems to stay relatively equally distributed after a higher usage in the early eighteenth century. For the genre analysis, the unbalanced representation of genres (a choice taken to represent the socio-historical context) hinders a valid diachronic analysis. In terms of differences in semantic relations, synonymy, antonymy, hyponymy and complementation haven been looked at, with the latter being the most prevalent semantic relation. The choice of excluding single occurrences due to the qualitative amount of data to be inspected lead to possibly excluding synonymy relations which the author expected to be much more prominent in the scientific discourse based on previous related work. On the other hand, the wide range of topics covered in CELiST is also considered a possible source of bias.

Chapter 7 analyzes female English scientific writing in CELiST. Botany writing has a large female tradition, which is clearly reflected in this corpus. The focus of the analysis is on directives as engagement features within scientific writing, that is, how the reader is engaged into the discourse. Female writers are compared to contemporary male writers. Knowledge about the surnames of each writer is essential to best comprehend Section 3. For the unknowledgeable reader, it is advisable to have the list of authors from pages 5–9 in Chapter 1 at hand, in order to know the gender of the author. A better audience design would have been advisable. An introductory section to the history of botany in the eighteenth and nineteenth centuries leads the ground to ask whether, in pragmatic terms, there is variation between female and male writers in how they render their discourse authoritative. For this, the use of directives is analyzed. The author well describes the challenges of finding directives properly in the corpus and how essential qualitative methods are to address this in a corpus-based fashion. Results show that male writers use directives more prominently than women in the eighteenth century and that in the nineteenth century women's engagement with the reader is

marked by delicate forms of engagement such as first-person plural combined with modal verbs.

The topic of female writers is continued in Chapter 8, where linguistic indicators of persuasion are analyzed. In focus are prefaces and dedications said to have a persuasive nature as evidenced in Section 2, which are then compared to main texts. The linguistic features analyzed are taken from the literature and encompass various forms of stance features. Again, the historical context is nicely introduced and valuable for any historical linguist. The results clearly show a prevalence of *to*-infinitives and first-person pronouns in prefaces by women writers as opposed to main texts. Wishful would have been a more qualitative analysis of the *to*-infinitive, as it is the most frequently used feature. In fact, from the few examples presented for the *to*-infinitive (cf. pp. 156–157) a tendency of its usage being an evaluative one seems to be quite evident, such as the *it-be-ADJ-to* pattern, a usual evaluative feature in academic writing, which increased its usage over time (cf. Hunston and Francis 2000 or Degaetano-Ortlieb 2015). It would have been interesting to see an evaluation of the possible different contexts the *to*-infinitive was used in.

Chapter 9 focuses on suasive verbs and compares CELiST with a corpus of non-fiction texts from the twentieth and twenty-first centuries. While comparatively this seems an odd selection, the diachronic insights gained taking this perspective are indeed valuable. The author shows how suasive verbs as a persuasion strategy are increasingly used in more contemporary non-fiction texts and seem to promote audience design in terms of the involvement of the reader. Moreover, women are more prominently using this persuasive strategy than men over time. A small typo in Figure 9 (*snd* to *and*) has found its way into the text.

Chapter 10 slightly changes the focus to the evolution of scientific practice within the scientific register, while accounting for the authorial presence of the author in the text through the analysis of conditionals and citation sequences. Yet, it still eludes at the evaluative character of the texts, specifically the authors stance towards the texts content. The author shows how epistemic evaluations (certainty-uncertainty) through conditionals and quotations are used within CELiST, and how these usages are complemented by attitudinal or stylistic ones.

The epistemic nature of the CELiST corpus is further investigated in Chapter 11, where epistemic adverbs are considered. The focus is on how writers of the life sciences

persuaded their readers to believe in the truth value of their statements. The comparison with the *Corpus of Historical English Texts* (CHET; cf. Moskowich *et al.* 2019) would have profited from either including evidence from the corpus or at least introducing reference to the respective work. This chapter also makes extensive use of the metadata provided in the CELiST corpus (authors' age, gender, time) showing how mid-career authors most frequently used epistemic adverbs to promote the truthfulness of their statements, how women underused them —possibly due to the descriptive register they were publishing for (botany)— and that their usage was most prominent in articles, lectures and essays.

Chapter 11 presents a very detailed analysis of *that* complement clauses in CELiST accounting for gender differences. Methodologically, the authors would have profited from a linguistically annotated version of CELiST such as part-of-speech annotation. Great effort is taken to extract *that* complement clauses relevant to the analysis. Detailed inspection of various variables combined with statistical evaluation allow the authors to draw valuable insights on the evaluative use of *that*-structures in CELiST. While the distributions provide evidence of no differences in use between female and male writing, by considering the evaluative functions and local contextual settings of evaluative *that* complement clauses the authors arrive at insightful conclusions towards a preference for a cognitive way of expression of scientific claims by female as opposed to a procedural way adopted by male writers. Some typos should be corrected (for instance, the word *attitudinal* is incorrectly written in the graphs on page 232 and *human-subjective* lacks a space on page 237).

Chapter 13 considers authority and deontic modals in CELiST. After a definition and a methodology section, a quantitative analysis of deontic modals follows. The quantitative analysis shows significant differences in the use of particular deontic modals between female and male, as well as regarding distributions across modal use. The qualitative section elaborates on the functions these modals fulfill, considering various functions which provide insights on the discursive patterns. It would have been nice to include a qualitative inspection here as well or, at least, explain why that might not have been possible.

Chapter 14 introduces a different aspect from evaluation and looks at coherence relations by the use of conjunctions in CELiST. It also uses the metadata provided in a fashionable way closely connected to the history of science of the field and categorizing

each metadata into meaningful categories (genres into specialized and non-specialized texts). The results show a steady increase in the use of the analyzed conjunctions over time. While all types of conjunctions rise, adversative and causal ones are definitely the most frequent ones. However, given the high frequency of *and*, which is excluded from the analysis, the reader might wonder whether the picture would have been different. The most important finding is that of a higher use of conjunctions in non-specialized texts as opposed to specialized ones, possibly favoring ease of processing in that genre. The argument of explicit mentioning is not quite straightforward as there is no comparison to implicit relations, so statements in this direction should be made with caution.

Chapter 15 is most quantitative in nature using multidimensional analysis to inspect register-internal variation of CELiST, including comparison to the *Corpus of English Texts on Astronomy* (CETA; cf. Moskowich and Crespo 2012) and the *Corpus of English Philosophy Texts* (CEPhiT; cf. Moskowich 2016). The focus is on the dimension of variation of descriptive and argumentative style. By comparison to the other disciplines, Life Science (CELiST) is fundamentally descriptive as opposed to Astronomy (CETA). Genre and gender differences round the picture nicely up also in light of the studies preceding this chapter.

Overall, the book is a great complement to the preceding series of the *Coruña Corpus of Early Scientific Writing*.² Two general remarks relate to (1) the cohesiveness of the single contributions and intended audience and (2) the contextualization of the studies to the international endeavors of historical corpus-based work as well as computational historical linguistics. As for (1), the chapters would have profited from more intersectional reference, especially because most of the analytical chapters engage in the topic of evaluative language. An introductory chapter to the books' single contributions would have been of great value to the interested reader. Here, the particular foci of the book could have been highlighted especially for the analytical chapters, such as the set of papers on evaluative language in CELiST as well as the more quantitative parts as opposed to more qualitative work. The gender aspect is taken up in the preface very nicely boosting interest in this direction. As for (2), while extensive related work has been considered in all contributions, the more international and more contemporary work of various historical corpus-based work has been rather

² <https://varieng.helsinki.fi/CoRD/corpora/Coruna/>

neglected. For example, the work around Tanja Säily's group³ (Helsinki) on English female writing in the Helsinki corpora of correspondences faces similar challenges in corpus building and seems relevant considering the contributions on female writing. The work by Elke Teich's group⁴ (Saarbrücken) on English scientific writing of the *Royal Society of London* provides a huge corpus and various linguistic annotations compared to the CELiST corpus, whose focus is on providing a qualitatively high resource truthful to the originals and authors' language. A smaller remark is directed at the poor linguistic annotation of the CELiST corpus, which would have profited from part-of-speech tagging or at least an explanation in the first chapters of why linguistic annotations have not been integrated. This said, the book presents a great endeavor taken to create the CELiST corpus, with a lot of effort put into beautifully enriching the corpus with valuable metadata and the aim to achieve a highly qualitative corpus resource reflecting the socio-historical setting of the time. Especially the transparency of the decisions made is to be highlighted.

REFERENCES

- Jenset, Gard B. and Barbara McGillivray. 2017. *Quantitative Historical Linguistics: A Corpus Framework*. Oxford: Oxford University Press.
- Hunston, Susan and Gill Francis. 2000. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.
- Degaetano-Ortlieb, Stefania. 2015. *Evaluative Meaning in Scientific Writing: Macro- and Micro-analytic Perspectives Using Data Mining*. Saarland: Saarland University dissertation.
- Moskowich, Isabel. 2016. Philosophers and scientists from the modern age: Compiling the *Corpus of English Philosophy Texts* (CEPhiT). In Isabel Moskowich Gonzalo Camiña Rioboó, Inés Lareo and Begoña Crespo eds. *'The Conditioned and the Unconditioned': Late Modern English Texts on Philosophy*. Amsterdam: John Benjamins, 1–23.
- Moskowich, Isabel and Begoña Crespo eds. 2012. *'Playne and simple': The Writing of Science between 1700 and 1900*. Amsterdam: John Benjamins.
- Moskowich, Isabel, Luis Puente-Castelo, Begoña Crespo and Leida María Monaco. 2019. *Writing History in Late Modern English. Explorations of the Coruña Corpus*. Amsterdam: John Benjamins.

³ <https://www2.helsinki.fi/en/researchgroups/varieng/corpus-of-early-english-correspondence>

⁴ <https://sfb1102.uni-saarland.de/projects/information-density-in-english-scientific-writing/>

Reviewed by

Stefania Degaetano-Ortlieb

Saarland University

Department of Language Science and Technology

Campus A2.2

66123. Saarbrücken

Germany

E-mail: s.degaetano@mx.uni-saarland.de

Review of Castro-Chao, Noelia. 2021. *Argument Structure in Flux: The Development of Impersonal Constructions in Middle and Early Modern English, with Special Reference to Verbs of Desire*. Bern: Peter Lang. ISBN: 978-3-034-34189-9.
<https://doi.org/10.3726/b17694>

Ayumi Miura
Osaka University / Japan

1. INTRODUCTION

To provide some background information about this volume and the author, which would have been included in the missing acknowledgement, the book under review is based on Noelia Castro-Chao's PhD thesis submitted to the University of Santiago de Compostela and defended in May 2020. Given that all the quoted online resources were last accessed before the end of 2019 and no publication from 2020 onwards is acknowledged, and judging from words like "[a]s of 2019" (p.80), the actual content of the book may be essentially identical to that of the thesis.

Anyone with serious interest in English historical syntax knows that there is a substantial amount of critical literature on impersonal constructions. Nevertheless, this book demonstrates that it is possible to make a new contribution. By examining the historical development of three formerly impersonal verbs of Desire after the general demise of impersonal constructions, Castro-Chao has successfully expanded the scope of research mainly in two respects: focus on the Early Modern English (EModE) period and a corpus-based investigation with *Early English Books Online Corpus 1.0* (EEBOCorp 1.0; Petré 2013). Most scholars have concentrated on Old English (OE) and/or Middle English (ME) because these are the periods when impersonal constructions are abundantly attested, and the transition to personal constructions is said to have been complete by the end of ME. EModE is often overlooked as the period of investigation,



despite its potential for offering insight into the situation immediately after the transition. In addition, previous studies tend to discuss examples extracted from dictionary entries or print editions. Extensive diachronic corpus-based research has therefore been lacking.

2. CHAPTER-BY-CHAPTER SUMMARY AND DISCUSSION

The book has nine chapters. The first five chapters set the context by describing the aims and the outline of the study (Chapter 1), reviewing previous studies (Chapter 2), introducing the theoretical framework for the research (Chapter 3), defining verbs of Desire (Chapter 4) and explaining the data and methodology (Chapter 5). Chapters 6, 7 and 8 form the main part of the book and are devoted to case studies of *lust*, *thirst* and *long* respectively. These three chapters have the same organisation: the first section concerns the origin and development of the verb in question based on the *Oxford English Dictionary* (OED), *Middle English Dictionary* (MED) and previous studies; the second section surveys the complements of (im)personal patterns historically recorded with the verb, based on the same sources; the third section looks at (im)personal patterns found in EEBOCorp 1.0 (1500–1700), with subsections for personal patterns arranged in decreasing order of frequency of complements; and the last section presents a summary and conclusions. Chapter 9 concludes the book.

2.1. Chapter 1: Introduction

The aims of the study are reproduced below (Section 1.1, see especially p.13):

- 1) To determine the time when the selected formerly impersonal verbs of Desire effectively ceased to be recorded with impersonal constructions.
- 2) To provide a diachronic overview of the personal syntactic patterns which came to replace impersonal constructions with these verbs from late ME onwards.
- 3) To describe the syntactic and semantic properties of the arguments of each individual verb studied.
- 4) To reflect upon factors which have been claimed to affect the loss of impersonal patterns in the history of English.
- 5) To assess which factors may have influenced the direction of the development of impersonal verbs of Desire after they started to appear in personal use.

Verbs of Desire constitute a syntactically coherent class in Present-day English (PDE; Levin 1993: 194–195), with some of them having alternated in OE and/or ME between

impersonal use with an objective Experiencer (e.g. *to þe me longeð swuðe* ‘I feel a great desire for you’) and personal use with a nominative Experiencer (e.g. *Ich langy so swiþe after Gorloys his wifue* ‘I have such a great desire for Gorloys’s wife’). The thing desired, Target of Emotion (ToE), is expressed today either as a direct object (*Dorothy needs new shoes*) or as a prepositional object (*Dana longs for a sunny day*). Castro-Chao duly justifies why EModE is a historically interesting period for impersonal constructions and how important it is to look closely at individual verbs. The ensuing chapters are outlined in Section 1.2.

2.2. Chapter 2: The function and development of English impersonal constructions

The chapter starts with a definition of ‘impersonal’ verbs and constructions (Section 2.1). Impersonal verbs are those predicates which occur in impersonal constructions though they may appear in other constructions and which subcategorise for an Experiencer as an obligatory argument, while personal verbs are restricted to personal constructions with a nominative Experiencer. Impersonal constructions lack a grammatical subject controlling verbal agreement, whereas personal constructions involve such a subject.

The next two sections introduce some of the crucial previous studies. Section 2.2 is concerned with those from the twentieth century, which all discuss the causes for the loss of impersonal constructions. The account begins with Jespersen’s (1961[1927]) well-known reanalysis hypothesis, which proposed that morphological ambiguity between a nominative noun phrase (NP) and a dative NP after the syncretism of these case forms primarily brought about the reanalysis of formerly impersonal constructions as personal constructions. This scenario has been criticised, especially on the grounds that impersonal constructions continued to be productive well after the simplification of the case system. Castro-Chao also draws attention to Allen’s (1986) study of OE/ME impersonal verb *like*, the very verb used in Jespersen’s hypothesis, which occurred with two nominal arguments only very infrequently. She then refers to alternative theories presented in Fischer and van der Leek (1983, 1987) and Allen (1986, 1995). Section 2.3 summarises more recent approaches which bear on the semantics of verbs and constructions: Trousdale (2008), Möhlig-Falke (2012) and Miura (2015).

Next (Section 2.4), the author presents a historical overview of impersonal constructions based on Möhlig-Falke (2012). As impersonal constructions disappeared

between 1400 and 1500, they were replaced with five syntactic alternatives: i) Experiencer-subject constructions (e.g. *She likes money*); ii) Experiencer-object constructions (e.g. *Her decision pleased me*); iii) (h)it-extraposition constructions (e.g. *It seemed to him that the weather would not last*); iv) middle-reflexive patterns, which are now obsolete (e.g. *They rate the goods without reason as they lust themselves*); and v) passive/adjectival patterns (e.g. *I am not quite pleased with your looks*).

The chapter finishes with a brief description of the semantic-pragmatic function of impersonal constructions (Section 2.5). According to Möhlig-Falke (2012), the OE impersonal construction expressed a shift of perspective by backgrounding/suppressing a nominative subject and foregrounding a dative/accusative Experiencer. It is considered functionally similar to the middle construction (Kemmer 1993) because they both involve an unvolitional event/process without any conceivable Causer.

2.3. Chapter 3: The nature of verb meaning and constructional meaning

Section 3.1 on verb meaning introduces key concepts such as State of Affairs (SoA) and semantic frame. The SoA of verbs of Desire is characterised by dynamicity, control and causation. Causation is further related to the force-dynamic relationship between an Initiator and an Endpoint or the causal chain which represents the transmission of force between the two participants. Then introduced is Dowty's (1991) concept of Proto-role and the Argument Selection Principle, which postulates that the argument with the largest number of Proto-agent properties is encoded as subject and the argument with the largest number of Proto-patient properties is encoded as direct object. Furthermore, in Dowty's Corollary 2, the non-subject argument with the largest number of Proto-patient properties is encoded as direct object, and the non-subject argument with the fewest Proto-patient properties is realised as an oblique or prepositional complement. Corollary 2 is concerned with three-place predicates (e.g. *John put the lamp on the table*), but Castro-Chao proposes extending it to two-place predicates with a prepositional phrase (e.g. *be afraid of NP*).

Section 3.2 addresses constructional meaning within Goldberg's (1995, 2006) model of Construction Grammar. The author carefully describes some terms and concepts pertinent to the data analysis in the book, such as participant/argument role and

(in)definite null complement. The section ends with a short discussion of perspective, whose relevance to impersonal constructions is proposed in Möhlig-Falke (2012).

Sections 3.3 and 3.4 deal with the semantic domains of Physical Sensation and Emotion respectively, which are both relevant for verbs of Desire. The semantic frame of verbs of Physical Sensation has the Feeler (Experiencer), the Body-part (Location) and the Cause (Stimulus). Verbs of Physical Sensation may denote either a process with a physical change of state (e.g. *Mary hurt John in the leg*) or a state (e.g. *My eyes are itching*), lack intention and control on the part of the Feeler and can be either causative or non-causative. In turn, verbs of Emotion typically involve an Experiencer and a Stimulus (more specifically a Cause or a ToE), encode a dynamic process between an Initiator and an Endpoint and exhibit a two-way causal relation where, on the one hand, the Experiencer as Initiator may direct their attention to the ToE (= unaffected Endpoint; e.g. *Mary likes John*), and on the other, the Cause as Initiator may bring about a mental state in the Experiencer (= affected Endpoint; e.g. *John pleases Mary*). This bidirectional relation allowed the same verb in early English to alternate between a nominative Experiencer and an accusative/dative Experiencer.

2.4. Chapter 4: The class of verbs of Desire

Castro-Chao first explains how she selected the three verbs to study (Section 4.1). She looked up the label ‘impersonal’ in the OED and MED entries of Levin’s (1993) twenty PDE verbs of Desire, and only four (*hunger*, *long*, *lust*, *thirst*) turned out to be documented in impersonal use in the definition of this book and in the sense ‘to desire’. All four verbs take a prepositional complement in PDE as members of Levin’s *long* verbs. *Hunger* is excluded from the investigation because it is very close to *thirst* in the development of the emotion sense and complementation patterns, as far as we can tell from their OED and MED entries. However, this reason does not justify why *thirst* deserves more attention than *hunger*. It is also left unexplained why *yearn* did not join the final list, except that the author may have dismissed its impersonal use in late ME, which is explicitly acknowledged in the MED entry (s.v. *yernen*), as a nonce expression.

Section 4.2 describes the semantic classification of verbs of Desire in the *Historical Thesaurus of the Oxford English Dictionary* (HTOED). ‘Desire’ (02.05.03.07) is a subfield of ‘Wish or inclination’ (02.05.03), comprised under ‘Will’ (02.05), one of the

seven subcategories of the major division ‘The mind’ (02). Not all of Levin’s twenty verbs of Desire are members of the HTOED category ‘Desire’, but the three verbs examined in the book all belong to this category.

Section 4.3 offers a semantic characterisation of verbs of Desire in the framework introduced in Chapter 3. The semantic frame of verbs of Desire has Desirer (Experiencer) and Desired (ToE). The SoA denoted by these verbs has a dynamic relationship, where the Desirer/Initiator directs their attention to the Desired/Endpoint. The SoA also possesses the feature of control in that the Desirer is volitional, but it lacks causation because the feeling of desire is not directly caused by the Desired. Moreover, the Desirer and the Desired differ only in the Proto-agent feature of volition and the Proto-patient feature of change of state. The Desired even lacks all the features of a prototypical Endpoint. Thus, verbs of Desire are semantically low in transitivity.

Finally, Section 4.4 outlines the syntactic patterns of PDE verbs of Desire according to Levin (1993). All these verbs occur in Experiencer-subject constructions. Some have developed adjectival patterns, specifically combination of a copula verb and either a past participle (e.g. *She was amazed/ashamed/disgusted/surprised*) or a related adjective (e.g. *He was afraid/angry/happy/hungry/sad/thirsty*), though only the latter seems to be relevant to verbs of Desire. These adjectival constructions usually have a stative interpretation and, just like impersonal constructions, background the Initiator and foreground the Endpoint of the SoA.

2.5. Chapter 5: Data and methodology

The chapter first presents basic information about *Early English Books Online* (EEBO; Davies 2017), its Text Creation Partnership (TCP) version and EEBOCorp 1.0, a 525-million-word corpus covering the period 1473–1700, from which Castro-Chao drew data for this study (Section 5.1). Next, Section 5.2 describes how she selected texts randomly from EEBOCorp 1.0 and compiled four fifty-year subcorpora, each with about five million words, for a diachronic study of the whole of EModE (1500–1700). Texts written at least partly in verse were excluded for fear of metrical interference on the choice of syntactic patterns. When the four subcorpora were finalised, the author used *AntConc* (Anthony 2019) to create a list of forms and spellings attested for the three verbs studied and to run a concordance search (Section 5.3). After false hits, repeated instances and

direct quotes of Biblical verses were all manually removed, *lust*, *thirst* and *long* respectively had 273, 304 and 341 valid examples. These instances were annotated according to ten variables (Section 5.4): i) subperiod of the corpus; ii) subject domain of the source text; iii) type of syntactic construction (impersonal or personal) and complements for personal patterns; iv) main or subordinate clause; v) type of subordinate clause; vi) formal realisation of the Desirer/Feeler; vii) formal realisation of the Desired; viii) type of preposition for the Desired; ix) person and number of the pronominal Desirer/Feeler; and x) Proto-role properties of the verb's participants.

2.6. Chapter 6: Lust

Lust originates in ME *lusten* (Section 6.1) and was found in impersonal constructions from the twelfth century to the mid-sixteenth century (Section 6.2.1). Impersonal constructions in subordinate clauses, especially those introduced by *as* or *when*, have a variant called NO PROP constructions, where the proposition is ellipped but easily recoverable (e.g. *Do as thee lust [to do] the terme of al thy lyf* ‘Do as it pleases you [to do] for the duration of all your life’). Personal patterns started to occur in the late fourteenth century (Section 6.2.2).

In EEBOCorp 1.0 (1500–1700), *lust* decreases remarkably in its overall frequency and is mostly confined to religious and Biblical contexts in each fifty-year subperiod (Section 6.3). The PDE specialised sense ‘to have a carnal desire’ gradually becomes more common as the general sense ‘to desire’ decreases. Impersonal constructions account for just about three per cent of all the tokens of *lust*. They all belong to the first subperiod (1500–1549) and are restricted to NO PROP constructions (Section 6.3.1).

Personal constructions have four complementation patterns (Section 6.3.2): patterns with clausal complements, patterns with zero complements, prepositional patterns and patterns with NP complements. Clausal complements are by far the most common choice in the first subperiod but sharply decrease over time, whereas zero complements show a parallel increase and overwhelm other patterns in the seventeenth century. Prepositional complements also become more prevalent, and NP complements are only sparsely attested in the first and the last (1650–1700) subperiods. Except for zero complements, all these patterns distinctly prefer pronominal Desirers.

Clausal complements in personal patterns have some NO PROPs taking the form of (SOV) fused relative constructions, where the referent of the subject of the main clause is given a free choice (e.g. *every man hath his fre will to doo what him lusteth* ‘every man has his free will to do what pleases him [to do]’) (Section 6.3.2.1). Castro-Chao proposes that patterns with clausal complements and patterns with (pro)nominal complements are functionally different, particularly in the Proto-patient property of independent existence. She then hypothesises that as clausal complements declined around the turn of the seventeenth century, and as semantic specialisation of *lust* proceeded, verbs which express more general meanings (e.g. *please, wish*) took over the functional space occupied by clausal complements.

Patterns with zero complements include examples with an adjunct prepositional phrase (PP), which account for almost 60 per cent of all the occurrences of zero complements (Section 6.3.2.2). These adjunct PPs are always headed by *against* or *contrary to*, and nearly 40 per cent of them are found in the fossilised expression *the flesh lusts against/contrary to the spirit* and its variants. Apart from this expression, which becomes more frequent over time, the overall frequency of zero complements is attributed to religious discourse, where the unexpressed object of desire tends to be implicitly assumed to refer to sin.

Prepositional patterns generally prefer a nominal Desired (Section 6.3.2.3), and patterns with NP complements invariably take a nominal Desired too (Section 6.3.2.4). Castro-Chao puts forward the hypothesis that patterns with NP complements arose in the sixteenth century as an alternative to prepositional patterns, only to go out of use by the end of the seventeenth century because the Desired lacks typical Proto-patient properties to maintain its status as NP object.

2.7. Chapter 7: Thirst

Thirst is of native origin and had two meanings in OE, ‘to feel thirst’ and ‘to desire’ (Section 7.1). The first physical-sensation sense encompasses a Feeler and a Needed in the semantic frame of the verb, though only the Feeler is lexically profiled. The sense ‘to desire’ is an extension from the physical-sensation sense and has the same semantic frame as *lust*, lexically profiling a Desirer and a Desired. Impersonal patterns with *thirst* are recorded from OE to the fifteenth century (Section 7.2.1). Personal patterns were already

available in OE but were less frequent than impersonal patterns and mainly occurred in texts with some Latin influence (Section 7.2.2).

During EModE, *thirst* decreases remarkably in its overall frequency and is mostly restricted to the religious domain (Section 7.3), just like *lust*. The sense ‘to desire’ is more common than the sense ‘to feel thirst’ in the early sixteenth century and gradually expands until the last subperiod (1650–1700), when the sense of thirst notably increases. EEBOCorp 1.0 (1500–1700) lacks instances of impersonal patterns, suggesting that the transition to personal constructions was virtually complete before 1500.

Personal patterns in the corpus are all Experiencer-subject constructions with either a prepositional complement, zero complement, NP complement or clausal complement (Section 7.3.1). All complementation patterns generally prefer pronominal Desirers. Only patterns with zero complements are regularly associated with the sense of thirst, and the other three patterns are all consistently connected with the sense of desire.

The Desired in prepositional patterns is mostly nominal (Section 7.3.1.1). Almost 30 per cent of Desirers are also nominal, and the author draws attention to body-part subjects like *heart* and *flesh*, which locate the emotion in a specific part of the body and cause an unvolitional interpretation of the SoA. The potential functional connection between body-part subjects and impersonal constructions leads Castro-Chao to hypothesise that body-part subjects were introduced to compensate for the meaning formerly expressed by impersonal constructions. She also discusses sentences where a drink noun occurs as part of the prepositional complement (*this drinke let vs thurst for*). She proposes to label this prepositional construction the MOVE-ATTENTION construction, where the Desirer moves or directs their attention to the Desired. The MOVE-ATTENTION construction is a metaphorical extension of the INTRANSITIVE MOTION construction (e.g. *The boy ran to the house*). Both constructions profile only the Experiencer/Theme argument and express the Target/Goal as an unprofiled oblique complement.

Patterns with zero complements were eventually replaced by the adjectival construction *to be thirsty* (Section 7.3.1.2). The data in EEBOCorp 1.0 (1500–1700) indeed show that *to be thirsty* steadily increases in frequency up to the third subperiod (1600–1649), when it overwhelms *thirst* with zero complement. Castro-Chao hypothesises that *to be thirsty* superseded zero complements with *thirst* because the stative interpretation of the adjectival construction and the semantic properties of the subject match the type of SoA expressed by *thirst* better than zero complements.

The Desired in patterns with NP complements is always nominal (Section 7.3.1.3). The author assumes that NP complements practically went out of use after the sixteenth century because the Desired lacks the Proto-patient properties for a prototypical object. Similarly to clausal complements of *lust*, clausal complements of *thirst* lack the Proto-patient property of independent existence and express an event in an unrealised future time (Section 7.3.1.4).

2.8. Chapter 8: Long

The native verb *long* lexically profiles Desirer and Desired (Section 8.1). It is recorded in impersonal patterns from OE until the first half of the sixteenth century (Section 8.2.1). Personal patterns began to appear in the early thirteenth century (Section 8.2.2).

Like *lust* and *thirst*, *long* decreases in frequency in the course of EModE and is often found in the religious domain (Section 8.3). EEBOCorp 1.0 (1500–1700) has no example of impersonal use, which implies that sixteenth-century instances in the dictionaries are only marginal. All the instances of personal patterns illustrate Experiencer-subject constructions, which take a prepositional complement, clausal complement, zero complement or adverbial complement (Section 8.3.1). Prepositional patterns are the most frequent in each subperiod and show a slight tendency to increase. In contrast, clausal complements gradually decrease while remaining the second most frequent pattern. The other two complements are very scarce, and NP complements are unattested. Desirers are generally pronominal irrespective of the type of formal realisation of the Desired.

The discussion of prepositional patterns includes patterns with adverbial complements because they share some functional properties (Section 8.3.1.1). All five instances of adverbial complements are realised by *therefore* ‘for that’. Castro-Chao suggests conceptualising the Desired *therefore* as a metaphorical Goal, just as *thereat* ‘to that place’ and *upward*, which are illustrated with *long* in the OED, express the Goal of literal directed motion. The construction with *therefore* is then considered as a variant of the above-mentioned MOVE-ATTENTION construction.

Patterns with clausal complements have sporadic instances of finite clauses introduced by *till/until* (e.g. *Christ longs till thou be in heaven*; Section 8.3.1.2). These clauses function as an equivalent of clauses with the declarative complementiser *that*

rather than their prototypical use as adverbial subordination of time. A number of adverbial subordinators (e.g. *as if/though*, *if*, *lest*, *like*) are known to have acquired the role of the major declarative complementiser *that* while retaining their original semantic features. Castro-Chao regards the use of *till/until* with *long* as another illustration of this development. It would be interesting to know how widespread the complementiser *till/until* was historically, especially in EModE, and if there are any generalisations to be made about the semantic categories of co-occurring verbs.

Zero complements with *long* parallel those with *lust* in that they involve definite null complements: the lexically profiled Desired argument is left unexpressed because it may be co(n)textually recoverable (Section 8.3.1.3). Therefore, *long* and *lust* are sometimes coordinated with each other in religious context where the unexpressed Desired is understood to refer to sin. In contrast, zero complements with *thirst* involve indefinite null complements which cannot be co(n)textually inferred. As a result, EEBOCorp 1.0 (1500–1700) has no example of *long* coordinated with *thirst* in patterns with zero complements.

2.9. Chapter 9: Discussion and conclusions

The first three sections (Section 9.1 to 9.3) address the first three aims of the study presented in Chapter 1 by summarising the findings for *lust*, *thirst* and *long* respectively. One noteworthy point in my view is that all three verbs decreased in overall frequency in the course of EModE. *Lust* may have become gradually infrequent as its semantic specialisation proceeded, and *thirst* with zero complements came to be replaced with *to be thirsty* especially in the sense ‘to feel thirst’. However, it is not clear from Chapter 8 or the summary in Section 9.3 why *long* decreased in frequency, if not as much as *lust* and *thirst*. It would be worth examining whether other verbs of Desire became more frequent in parallel or if there was any large-scale shift in conceptualisation of desire in EModE.

The next two sections deal with the last two aims of the study. The author’s findings confirm some of the previous accounts about the disappearance of impersonal constructions (Section 9.4). On the other hand, the loss of case distinctions and the rigidification of word order –the two morphosyntactic changes which are often quoted in the literature to explain the demise of impersonal constructions– do not necessarily match

the period of transition from impersonal to personal constructions with the three verbs in question. Castro-Chao's most crucial counterargument is against Jespersen's (1961[1927]) reanalysis hypothesis. She adds to criticisms in previous studies with her EModE data, where Desirers are mostly realised as pronouns for all three verbs. Just as OE/ME *like* governed two nominal arguments highly infrequently (Allen 1986), Jespersen's syntactic scenario with two morphologically ambiguous nominal arguments was not frequent with EModE *lust*, *thirst* and *long* either.

Section 9.5 addresses questions as to why verbs of Desire adopted Experiencer-subject constructions and why they developed prepositional patterns rather than patterns with NP complements. The corpus data showed that Desirers are predominantly pronominal while Desired is more likely to be nominal and can also be a PP or a clausal complement. The principle of end-weight inevitably placed the lighter Desirer in the preverbal position, and when SVO word order was fixed, the Desirer became the best candidate for subject. In addition, the Desirer shows the Proto-agent feature of volition, which is eligible for a subject rather than an object. This may be why *thirst* and *long* never occur in Experiencer-object constructions or (*h*)*it*-extraposition constructions. However, *lust* is attested in Experiencer-object constructions (SOV fused relative constructions; see Section 6.3.2.1) in EEBOCorp 1.0 (1500–1700) and (*h*)*it*-extraposition constructions during ME. This issue is left for further research. The answer to the second question above concerns interaction between verb meaning and constructional meaning: since verbs of Desire are apt to be conceptualised as a metaphorical inclination/movement towards something, they came to be associated with prepositional patterns, that is, the MOVE-ATTENTION construction, a metaphorical extension of the INTRANSITIVE MOTION construction which takes a prepositional Goal. The Desired in the MOVE-ATTENTION construction is an unaffected Endpoint without the majority of Proto-patient properties and suitable to be realised as a prepositional complement rather than a direct object, following Dowty's (1991) Corollary 2 (see Section 3.1). Consequently, *long* never takes a NP complement in EEBOCorp 1.0 (1500–1700), and *lust* and *thirst* cease to do so in the course of the sixteenth century. Whilst Castro-Chao's scenario sounds intriguing, Levin's (1993) PDE verbs of Desire consist not only of *long* verbs which indeed take a prepositional complement but also transitive *want* verbs (e.g. *covet*, *desire*, *fancy*, *need*, *want*). It is vital to investigate why verbs of Desire have these two subclasses or if there

are any subtle semantic distinctions between them which cause different complementation patterns.

A comparison between *long* verbs and *want* verbs is suggested in Section 9.6 as a topic for further research. Other ideas include extending the chronological coverage to late ME, enlarging the corpus especially in terms of genre, studying other impersonal and non-impersonal verbs of Desire such as *desire* and *hunger* and conducting a corpus-based study of the MOVE-ATTENTION construction in ME and EModE.

3. FINAL OBSERVATIONS

The book is based firmly on previous studies, most notably Möhlig-Falke (2012), and presents a fine piece of qualitative and quantitative research with rich empirical evidence and theoretical contribution. Furthermore, Castro-Chao has demonstrated that the historical development of these verbs is not just about the loss of impersonal patterns and shift to personal patterns. It bears on a number of other aspects such as the rivalry among different complements, Proto-role properties of verbal arguments, the principle of end-weight and information structure, the MOVE-ATTENTION and INTRANSITIVE MOTION constructions, the shift from adverbial subordinators to complementisers, among others. The author has thus successfully widened the scope of research about impersonal constructions. Other remarkable benefits include detailed descriptions which allow readers unfamiliar with impersonal constructions and all the tools/theories employed to follow the account effortlessly and confirm each crucial step in research, such as how the three verbs in question were determined and how data from EEBOCorp 1.0 (1500–1700) were collected. Finally, the author provides meticulous in-text acknowledgement and cross-references throughout the book and writes in readable English with accurate grammar and idioms. Obvious typos are only very sporadic, though there are some regrettable typesetting errors.

I entirely agree with Castro-Chao about suggestions for further research (Section 9.6). I even think some of them could have been attempted in this book. Focus on EModE is one of the unique features of this work. Nevertheless, given the objectives of the whole research cited in Chapter 1 (especially the first two), a close investigation into the last few decades of late ME would have reinforced the conclusion about the shift from impersonal to personal constructions and reflected *Middle and Early Modern English* in the title of

the book more accurately. EEBOCorp 1.0 starts from 1473, so even though it is impossible to create a fifty-year subperiod corpus before 1500, we would have obtained some data for comparison. A discussion of at least one near-synonymous non-impersonal verb of Desire would also have been appreciated. After all, only three verbs were examined, though thoroughly, and all of them are rather infrequent. Their development may have been affected by rivalry with more frequent near-synonyms which were never used impersonally.

Notwithstanding these minor criticisms and desiderata, I believe that Castro-Chao has produced a substantial work which will be essential reading for those interested in impersonal constructions, diachronic lexical semantics, diachronic Construction Grammar and corpus-based syntactic research.

REFERENCES

- Allen, Cynthia L. 1986. Reconsidering the history of *like*. *Journal of Linguistics* 22/2: 375–409.
- Allen, Cynthia L. 1995. *Case Marking and Reanalysis: Grammatical Relations from Old to Early Modern English*. Oxford: Clarendon Press.
- Anthony, Laurence. 2019. *AntConc* (version 3.5.8) [Computer Software]. Tokyo: Waseda University. <https://www.laurenceanthony.net/software/antconc/>
- Davies, Mark. 2017. *Early English Books Online*. SAMUELS project. <https://www.english-corpora.org/eebo/>
- Dowty, David. 1991. Thematic proto-roles and argument selection. *Language* 67/3: 547–619.
- Fischer, Olga and Frederike van der Leek. 1983. The demise of the Old English impersonal construction. *Journal of Linguistics* 19/2: 337–368.
- Fischer, Olga and Frederike van der Leek. 1987. A ‘case’ for the Old English impersonal. In Willem F. Koopman, Frederike van der Leek, Olga Fischer and Roger Eaton eds. *Explanation and Linguistic Change*. Amsterdam: John Benjamins, 79–120.
- Goldberg, Adele E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.
- Goldberg, Adele E. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.
- Historical Thesaurus of the Oxford English Dictionary*. 2009. Christian Kay, Jane Roberts, Michael Samuels and Irené Wotherspoon eds. <https://www.oed.com/thesaurus>
- Jespersen, Otto. 1961[1927]. *A Modern English Grammar on Historical Principles. Part III: Syntax (Second Volume)*. Copenhagen: Ejnar Munksgaard. [Reprinted, London: George Allen & Unwin].
- Kemmer, Suzanne. 1993. *The Middle Voice*. Amsterdam: John Benjamins.
- Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.
- Middle English Dictionary*. 1952–2001. Hans Kurath, Sherman M. Kuhn and Robert E. Lewis eds. Ann Arbor: University of Michigan Press. Online edition available at

- the Middle English Compendium. 2000–2018. Frances McSparran *et al.* eds. Ann Arbor: University of Michigan Library. <https://quod.lib.umich.edu/m/middle-english-dictionary/dictionary>
- Miura, Ayumi. 2015. *Middle English Verbs of Emotion and Impersonal Constructions: Verb Meaning and Syntax in Diachrony*. Oxford: Oxford University Press.
- Möhlig-Falke, Ruth. 2012. *The Early English Impersonal Construction: An Analysis of Verbal and Constructional Meaning*. Oxford: Oxford University Press.
- Oxford English Dictionary Online*. <https://www.oed.com>
- Petré, Peter. 2013. *Early English Books Online Corpus 1.0*. Leuven: KU Leuven.
- Trousdale, Graeme. 2008. Words and constructions in grammaticalization: The end of the English impersonal construction. In Susan M. Fitzmaurice and Donka Minkova eds. *Studies in the History of the English Language IV: Empirical and Analytical Advances in the Study of English Language Change*. Berlin: Mouton de Gruyter, 301–326.

Reviewed by

Ayumi Miura

Osaka University

Graduate School of Humanities

1–8 Machikaneyama, Toyonaka

Osaka, 560-0043

Japan

e-mail: ayumi.miura.hmt@osaka-u.ac.jp

Review of Wallis, Sean. 2020. *Statistics in Corpus Linguistics: A New Approach*. London: Routledge. ISBN: 978-1-138-58938-4. DOI: <https://doi.org/10.4324/9780429491696>

Tove Larsson
Northern Arizona University / United States

In this book, Sean Wallis provides an introduction to statistics as it applies to corpus linguistics studies. The book offers a valuable complement to existing resources available to researchers in the field, (i) by focusing on distributions and confidence intervals and (ii) by offering in-depth explanations of their underlying mathematics and logic. As stated in the preface, “what seems missing [in traditional books on statistics for corpus linguistics] is a clear explanation as to how a test procedure works from ground up” (p. xiii). The book has 19 chapters divided into six sections: 1. “Motivation,” 2. “Designing experiments with corpora,” 3. “Confidence intervals and significance tests,” 4. “Effect sizes and meta-tests,” 5. “Statistical solutions for corpus samples,” and 6. “Concluding remarks.” This review will follow the same outline and end with a brief evaluation of the volume.

In the first section, Wallis lays the foundations for the subsequent chapters by providing an overview of the field of corpus linguistics, different kinds of corpora, and study designs that researchers who work with corpora may use. He introduces three distinct classes of empirical evidence that can be obtained from a corpus: factual evidence of a linguistic event (i.e., a linguistic token), frequency of a linguistic event, and interaction evidence between two or more linguistic events (i.e., the probability that an event x will occur, given an event y). After introducing what is referred to as the ‘3A Cycle’ (Annotation, Abstraction, and Analysis), which is central to all corpus linguistic studies, Wallis subsequently addresses the question of what (richly) annotated corpora can tell us. The author also explicitly argues against a “simplistic ‘bigger is best’



approach” (p. 3) to data analysis and corpus building. The section concludes with some example studies and a discussion of framing constraints in study design.

The second section discusses how the scientific method is applied in corpus linguistic studies. It begins with an overview of the research process, including the compilation of a corpus (i.e., a sample), formulation of research questions and hypotheses, and evaluating these hypotheses through experiments and statistical tests. In the process, the author introduces concepts such as variables (and the fact that the variables we use are of different numerical types or scales: binomial, multinomial, ordinal, interval, and real). The author then widens the discussion to variationist designs, that is, designs where linguists study “the influence on [decisions that speakers or writers make when forming utterances], identifying factors that affect the selection of one option over another” (p. 47). Binomial and multinomial techniques are introduced and illustrated using linguistic examples. The author argues that such designs are preferable to designs that rely on word-based baselines (e.g. per-million-word frequencies), as “language is not a sequence of random words” (p. 48) and as such baselines do not “distinguish opportunity and choice, and are vulnerable to arbitrary variation” (p. 74). Finally, sampling as it applies to corpus linguistics is discussed. Example studies are used to illustrate the techniques and points made throughout.

Section 3 is devoted to inferential statistics with a specific focus on distributions and confidence intervals. Wallis builds on the discussion in Sections 2 and 3 to introduce foundational concepts such as *p*-values, distributions, and confidence intervals. Using the Wilson score interval and the Newcombe-Wilson Interval, the notion of ‘significant differences’ (in the null hypothesis statistical testing framework) is introduced. This section also includes a chapter on replication and ‘the replication crisis’ (which started with the observation that findings from many studies are difficult to reproduce). The author brings up possible reasons why findings in corpus linguistics studies may not replicate, such as differences across studies with regard to populations, samples, and/or operationalizations of key categories and constructs. He further gives recommendations for the field moving forward, including a checklist for empirical linguistics studies that emphasizes the need for accuracy and transparency in our reporting practices. The final chapter in this section deals with the question of how to choose the right statistical test depending on the type and scale of the variables of interest.

In Section 4, Wallis covers a discussion of effect size measures and meta-tests. First, measures of interdependence (i.e., measures of the bidirectional association between independent and dependent variables) such as Cramér's Φ are discussed. The author then moves on to introduce tests that can be used to compare results across studies: so-called 'meta-tests'. As pointed out by the author, such tests are helpful in the context of replication in that researchers may wish to compare results from (a) studies for which the data are kept constant, but where the design has been changed slightly, or (b) studies for which the design is kept constant, but where the data are different. Specifically, different Wilson-based tests are introduced and exemplified using linguistic data.

Section 5 discusses different statistical solutions for corpus samples. Wallis begins this section by stressing the importance of being able to justify the frequencies obtained from a corpus analysis. That is, we should not take output from taggers and concordancers at face value, but rather always assess and try to improve the accuracy of the annotation of the phenomenon of interest, especially for larger corpora that have not been manually checked. The author goes on to propose a golden rule of data: "We need to know that, as far as possible, our dataset is a sound and complete set of examples of the linguistic phenomenon in which we are interested" (p. 263). The remainder of the section is devoted to a discussion of how to recalibrate tests such as binomial models to account for sampling issues common to corpus linguistics (such as the fact that in many cases, the assumption of case independence is violated).

The sixth and final section of the book contains two chapters. The first includes an in-depth description of Wilson distributions (previously discussed in the volume in the context of the Wilson score interval) with the purpose of having the reader "understand the performance of the Wilson formula, distribution and interval itself" (p. 297). The impact of the size of the sample is also discussed. The final chapter of the book offers concluding remarks where the author comments on the content of the book and what he would like to see for the field moving forward. For example, the importance of making sure we have a reliable sample/data source is stressed: "if our source data are not what we think they are, all statistical generalisation must be in vain!" (p. 315).

From an evaluative perspective, this volume has a number of considerable strengths. First and foremost is the fact that the book is written specifically for a corpus linguistic audience using example studies and data from the field. As anyone who has taken statistics courses in other field knows, learning about a new technique using data

and study designs that are unfamiliar adds a layer of difficulty to a subject that can already be challenging, requiring the learner to ‘translate’ the new techniques between fields and figure out how they apply to our kinds of research questions. It is thus immensely helpful to have a resource such as the present volume that is written specifically with our kinds of data and analyses in mind.

Another clear strength of the volume, one in line with its stated goal of making inferential statistics more tangible to help readers understand what we are doing at all stages of the analysis, is its detailed illustrations of techniques and processes. The graphs and sample studies that are spread out across the sections are very helpful for making the statistical reasoning more concrete. In this context, Wallis makes important methodological points about the importance of rigorous study design (including sampling, corpus annotation, and analysis), as “statistical methods do not turn a poor experimental design into a good one” (p. 316). He even goes so far as to say that “significance testing is secondary to the primary task of plotting data and engaging with a linguistic evaluation of what our results might mean” (pp. 314–315). Because statistics, by some, may be perceived as necessarily involving an element of ‘black-box-iness’, this is a refreshing perspective, one where the main focus is on the linguistic analysis and where the statistical analysis is viewed as a tool that enables generalizations of the results of that analysis beyond a specific sample.

The subtitle of the book, *A New Approach*, is indeed apt in that, unlike previous books that focus more on traditional significance testing techniques, readers of this book get to view statistics primarily through the lens of distributions and confidence intervals. While this approach has many advantages (some of which are outlined above), one downside is that the book does not cover many of the techniques that researchers in the field would encounter in publications and courses and that they may therefore benefit from learning more about. To be fair, as stated in the preface, Wallis’s book is intended to “supplement, not replace other textbooks in statistics or linguistics” (p. xiv). As such, the ideal audience for the book may thus not be complete beginners in either corpus linguistics or statistics, but rather, perhaps, readers who already have some foundation in both and who wish to improve and complement their statistical reasoning and understanding in the context of corpus linguistics. Regardless of their level, however, it would perhaps be helpful for readers if future editions of the book could include some additional instructional materials (maybe as online supplements), such as exercises and

code for one or several software packages, to help readers get started applying their newfound knowledge.

Further, and related to the previous point, one of the main strengths of the books is, perhaps, also a limitation: the book covers a lot of ground in relatively few pages (approximately 350 pages). That is, while organized in a fairly logical manner, the book spans an impressive set of topics, both corpus linguistic and statistical, which inevitably means that some topics must be covered in more detail than others. At times, this requires readers to have fairly advanced knowledge of topics to be able to follow the line of argumentation. Some more sign-posting, interim summaries, and suggestions for further reading would have been helpful here.

All in all, Wallis offers a very interesting and —to corpus linguistics— new perspective on statistical analysis. In addition, the many points the author makes throughout the volume on the importance of transparency and rigor in our study designs are all well taken and tremendously important for a field that finds itself growing —and needing to grow— quantitatively. There is no doubt that this volume constitutes a very valuable resource for current students and researchers in corpus linguistics, one that will no doubt also continue to grow in future editions to meet our field's exciting future.

Reviewed by

Tove Larsson
Northern Arizona University
English Department
Box 6032
Flagstaff, AZ 86001
United States
e-mail: Tove.Larsson@nau.edu