

RiCL

**Research in
Corpus Linguistics**



RiCL 13/2 (2025)



aelinco

Asociación Española de Lingüística de Corpus

RiCL 13/2 (2025)

Editors

Zeltia Blanco-Suárez and Paula Rodríguez-Abruñeiras

ISSN 2243-4712

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

RiCL

Research in
Corpus Linguistics



Official journal of

aelinco

Asociación Española de Lingüística de Corpus

<i>Articles</i>	<i>Pages</i>
Exploring noun lexical diversity and noun phrase complexity in Spanish email writing at B1 and C1 levels Natalia Judith Laso Martín, María Belén Díez-Bedmar	1–34
The creation of the Indonesian TreeTagger for use in LanksBox and CQPweb Prihantoro	35–62
The <i>Multi-Feature Tagger of English</i> (MFTE): Rationale, description and evaluation Elen Le Foll, Muhammad Shakir	63–93
Same, same, but <i>erm sort of</i> different? Comparing three kinds of fluencemes across Australian, British, Canadian, and New Zealand English Karola Schmidt, Sandra Götz, Katja Jäschke, Stefan Th. Gries	94–123
The <i>Construction Complexity Calculator</i> (ConPlex): A tool for calculating Nelson’s (2024) construction-based complexity measure Christopher R. Cooper	124–143
The language of evaluation and stance in crowdfunding project proposals Alberto A. Vela-Rodrigo	144–174
Spanish EFL learners’ use of contrastive linking adverbials across three CEFR levels and gender Carmen Maíz-Arévalo	175–199
A corpus-based study on the transitive uses of English physiological verbs Beatriz Rodríguez Arrizabalaga	200–231
 <i>Book Reviews</i>	
Martín Arista, Javier and Ana Elvira Ojanguren López. 2024. Structuring Lexical Data and Digitising Dictionaries: Grammatical Theory, Language Processing and Databases in Historical Linguistics. Leiden: Brill. 412 pp. ISBN: 978-90-04-70266-0. https://doi.org/10.1163/9789004702660 Silvia Saporta Tarazona	232–240
Review of Ljubica Leone. Composite Predicates in Late Modern English (<i>Routledge Focus on Linguistics</i>). 2024. London: Routledge. 92 pp. ISBN 9781003410256. https://doi.org/10.4324/9781003410256 Harumi Tanabe	241–248

Exploring noun lexical diversity and noun phrase complexity in Spanish email writing at B1 and C1 levels

Natalia Judith Laso Martín^a – María Belén Díez-Bedmar^b
University of Barcelona^a / Spain
University of Jaén^b / Spain

Abstract – Research on noun phrase use in EFL writing has mainly focused on linguistic complexity and accuracy, lexical richness, and phraseological competence. However, the relationship between noun lexical diversity of nouns and the syntactic complexity of the noun phrases in which these nouns appear remains underexplored. To address this gap, this paper examines the lexical diversity of head nouns in noun phrases within a sample of emails written by L1 Spanish EFL learners at B1 and C1 proficiency levels, taken from the *FineDesc Learner Corpus*. The analysis considers both the lexical diversity of nouns and the syntactic complexity of the noun phrases they head. The findings reveal: a) a narrower range of nouns at the B1 level compared to the C1 level; b) a low percentage of nouns from both levels, based on the *English Vocabulary Profile*; and c) differences in NP complexity between the two proficiency levels (B1 and C1), depending on whether the head nouns are concrete or abstract. The paper underscores the importance of combining different complexity measures —namely, lexical diversity and NP complexity analyses— to gain a more comprehensive understanding of learners’ use of noun phrases.

Keywords – lexical diversity; NP complexity; learner email writing; CEFR B1; CEFR C1

1. INTRODUCTION¹

Much of the past and current literature on the noun phrase (NP) in learner language writing focuses on linguistic complexity and accuracy (Ortega 2003; Biber *et al.* 2011; Lu 2011; Bulté and Housen 2012, 2014; Ai and Lu 2013; Crossley and McNamara 2014; Parkinson and Musgrave 2014; Liu and Li 2016; Xu 2019; Díez-Bedmar and Pérez-Paredes 2020; Kim 2021), as well as lexical richness and phraseological competence (Howarth 1998; Biber and Conrad 1999; Nation 2001; Hyland 2008; Šišková 2012; Peters 2016; Vedder and Benigno 2016; Paquot 2019; Du *et al.* 2022). Additionally, the linguistic characteristics of EFL writing have been widely discussed in relation to learners’ L1, topic and genre effects, task complexity, and learners’ L2 level of

¹ This paper was supported by Grant PID2020-117041GA-I00, funded by MICIU/AEI/10.13039/501100011033. The authors would like to thank Arturo Montejo Ráez for his valuable assistance in the automatic annotation of nouns in our study.



proficiency (Ellis and Yuan 2004; Ong and Zhang 2010; Díez-Bedmar 2015; Mazgutova and Kormos 2015; Liu and Li 2016; Yoon 2017; Ionin and Díez-Bedmar 2021, among others). However, few studies have approached the analysis of learner language by combining syntactic and lexical complexity measures (see Section 2.3), which prevents a more comprehensive understanding of NP learner production. The present study contributes to the literature by considering two complexity measures, noun lexical diversity and NP syntactic complexity, to study NP production in email L1 Spanish learner writing at different CEFR levels. The levels selected are B1 and C1 and the text type selected is email writing.

The selection of B1 and C1 proficiency levels is driven by our objective to explore the differences between the first level of the independent user (B1_Threshold) and the first level of the proficient user (C1_Effective Operational Proficiency) as defined by the *Common European Framework of Reference for Languages* (CEFR; Council of Europe 2001). By comparing these two levels, we aim to gain insights into the lexical and syntactic development from an intermediate to an advanced stage of language proficiency.

The choice of email writing is justified by the fact that lexical and syntactic complexity has primarily been analysed in academic language, namely essays, (Šišková 2012; Treffers-Daller *et al.* 2018; Clavel-Arroitia and Pennock-Speck 2021; Lahuerta 2024, among others) rather than in transactional language. Transactional texts, such as email writing, are not only a prevalent text type in EFL writing but also hold significant importance in high-stakes language accreditation exams. Therefore, this focus allows us to contribute novel insights into an underexplored text type that is highly relevant for EFL learners.

In this study, the classification of nouns into different semantic categories (e.g., concrete vs. abstract nouns, hyponyms within specific semantic fields) is essential for understanding the relationship between noun choice and syntactic complexity. Nouns are central to the structure of NPs, and their semantic properties can influence NP complexity. Moreover, the use of hyponyms provides insight into lexical diversity in learner writing. By analysing noun types and their semantic classifications, we can understand better how learners' lexical choices at different proficiency levels impact the complexity of the NPs they produce. This classification is directly tied to our research questions, as it allows us

to explore whether more abstract or concrete nouns (or nouns from specific semantic fields) are associated with particular patterns of NP complexity at different CEFR levels.

To reach a better understanding of NP learner production by considering two complexity measures, noun lexical diversity and NP complexity types, in email writing at two CEFR levels, our study addresses the following two research questions:

RQ1: How does the type of noun hyponyms used in Spanish EFL email writing production correlate with the *English Vocabulary Profile* (EVP), according to CEFR proficiency levels?

RQ2: Does the semantic field of the head noun employed affect the NP complexity types produced in email writing by L1 Spanish learners of English at B1 and C1 levels?

In the next section we will examine existing research on NP complexity and lexical diversity, focusing on how these factors relate to CEFR levels and how they support the present study's research questions.

2. LITERATURE REVIEW

2.1. *Lexical diversity of nouns in learner writing*

Research has shown that lexical diversity —understood as the range and variety of words used in a text, and, more specifically, lexical variation, focused on lexical words— are key indicators of language proficiency, particularly in EFL contexts (Engber 1995; Crossley *et al.* 2011; Kuiken *et al.* 2010; Housen *et al.* 2011; Lu 2012; Vidal and Jarvis 2020; Allaw 2021). The exploration of lexical richness within the framework of proficiency levels defined by the CEFR has gained significant attention in recent learner corpus-based studies. Šišková (2012), for instance, conducted a study on lexical richness in narratives written by Czech EFL learners. Focusing on various measures of lexical diversity, the research revealed strong correlations between various indicators of lexical richness and underscored the importance of vocabulary knowledge in language acquisition. In a related study, Gregori-Signes and Clavel-Arroitia (2015) explored both lexical density and lexical diversity across different writing tasks produced by university EFL learners at B1 and C2 levels. Their examination of writing tasks revealed a progression in lexical diversity across proficiency levels. This progression of lexical

diversity B1-C2 highlights the importance of learner corpus research cross-sectional studies in assessing proficiency levels by examining lexical diversity in learner corpora. In a similar vein, other studies, such as Treffers-Daller *et al.*'s (2018), contribute to ongoing discussions on how lexical diversity measures can assist in differentiating essays by proficiency level. Their study revealed that basic measures of lexical diversity, such as the number of different words and type-token ratio (TTR), exhibited strong predictive power in discriminating between proficiency levels when controlling for text length. This finding underscores the usefulness of fundamental lexical indices in assessing language proficiency accurately.

Expanding on the investigation of lexical complexity, Su *et al.* (2023) conducted a nuanced analysis of exemplar EFL texts across different grade levels in China. They identified specific lexical diversity and sophistication features as effective markers of lexical complexity and explored their implications for assessing language proficiency and guiding text adaptation practices. Furthermore, Clavel-Arroitia and Pennock-Speck (2021) compared lexical density, diversity, and sophistication in written and spoken interactions of university students during English as a lingua franca telecollaborative exchange. While Spanish learners exhibited higher lexical diversity and sophistication in their written production, the differences in the oral production were subtler, suggesting context-dependent variations in lexical usage.

In addition to corpus-based investigations, theoretical frameworks have been developed to model lexical proficiency accurately. Crossley *et al.* (2011) undertook a thorough examination of lexical diversity in written texts, highlighting its close correlation with students' language proficiency levels. They proposed a model of lexical proficiency based on computational indices, including those related to noun usage. Their findings indicate that the number of different words, word hypernymy values, and content word frequency in a text gradually increases across different proficiency levels, reflecting thus a gradual increase in lexical diversity. The observation by Crossley *et al.* (2011) aligns with Qin and Uccelli's (2020) study, where they employed multi-level linear models to analyse lexical diversity as a dependent variable to capture differences associated with English proficiency. They found that lexical diversity increases throughout the language learning process, and it is often associated with more diverse vocabulary in academic writing.

Despite the wealth of research on lexical diversity and its implications for language proficiency in EFL writing, there remains a gap in understanding the lexical diversity of nouns within the CEFR framework in Spanish EFL written production. Therefore, an in-depth analysis of the lexical diversity of nouns in a learner corpus of Spanish EFL learners' email writing at B1 and C1 levels can enhance our understanding of the development of linguistic proficiency in EFL writing.

2.2. NP complexity in learner writing

Skehan's (1989) three-part model of L2 proficiency, which considers complexity, accuracy and fluency (CAF), has shaped the research conducted to analyse (learner) language production. Specially since Wolfe-Quintero *et al.*'s (1998) and Ortega's (2003) research syntheses, a plethora of publications have employed CAF measures to analyse texts produced by speakers in the target variety (e.g., English as an L1) and by speakers from other varieties, such as the learner varieties (e.g., Spanish learners of English).

The most frequently analysed type of complexity is syntactic complexity, which is defined as "the range and the sophistication of grammatical resources exhibited in language production" (Ortega 2003: 82), "the progressively more elaborate language that may be used, as well as a greater variety of syntactic patterning" (Foster and Skehan 1996: 303), or "a wide variety of both basic and sophisticated structures" (Wolfe-Quintero *et al.* 1998: 69). Therefore, the variety of syntactic forms produced, their sophistication, and degree of elaboration have been considered in the analyses of syntactic complexity.

The study of syntactic complexity at NP phrase level has been advocated for so that language complexification at different levels can be captured, thus allowing for the description of the multidimensional nature of the syntactic complexity construct (Norris and Ortega 2009; Biber *et al.* 2011; Lu 2011; Kyle and Crossley 2018; Casal and Lee 2019; Lan *et al.* 2022; Zhang and Lu 2022). The analyses have been conducted by using a variety of measures which, in different ways, have taken into consideration the constituents in the NP. Some studies have drawn from Biber *et al.*'s 2011 syntactic developmental index (e.g., Parkinson and Musgrave 2014; Staples *et al.* 2016; Ansarifard *et al.* 2018; Casal and Lee 2019; Lan *et al.* 2019). Other publications have operationalised NP complexity by considering the number of constituents in the NP with length measures (see Ravid and Berman 2010; Kuiken and Vedder 2019; Lu and Wu 2022, among others).

Other studies have considered the so-called ‘complex nominal’, even though differences in the operationalisation of such measure are found. For instance, the complex nominal is defined by Lu (2010: 483) as a noun which is modified by means of an attributive adjective, possessive noun, post-preposition, relative clause, participle or appositive, a noun clause, as well as gerund and infinitival subjects. However, Vyatkina (2013) considers within complex nominal structures attributive adjective phrases, prepositional phrases extending nominal phrases, nominal clauses, and relative clauses. Caution is therefore to be taken when comparing the results obtained in the different studies on the topic, as NP complexity is operationalised in different ways.²

Taking as a starting point the syntactic complexity measures in Wolfe-Quintero *et al.*’s (1998) and Ortega’s (2003), software such as the *L2 Syntactic Complexity Analyzer* (L2SCA; Lu 2010) and the *Tool for the Automated Analysis of Syntactic Sophistication and Complexity* (TAASC; Kyle 2016) have been developed to automatically analyse a number of syntactic complexity measures at sentential, clausal, and phrasal levels. Consequently, some publications have combined both, the automatic analysis, and the manual analysis to aim for a comprehensive analysis of NP complexity. For instance, Díez-Bedmar and Pérez-Paredes (2020) employed the nominal measures in TAASC and conducted a manual parsing of the syntactic complexity of all NP types, which revealed 29 different NP types divided into simple NPs (i.e., determiner NPs), premodified NPs, postmodified NPs, and pre- and postmodified NPs. The results showed that a combination of both analyses provide the more exhaustive results in the analysis of NP complexity in learner language, as NP is operationalised in different ways and a more fine-grained analysis may be obtained.

Despite the wealth of studies on NP complexity, there is little information on NP complexity considering different CEFR levels. To the best of our knowledge, most of the studies available do not focus on the NP exclusively, but measures related to the NP are found together with other syntactic complexity measures. This is the case of Khushik and Huhta (2020), who offered information regarding complex nominals per clause and complex nominal per T-unit,³ and modifiers per noun phrase and noun phrase density,

² Different labels are used in the literature to refer to the phrase whose head is a noun or a pronoun. The umbrella term is ‘noun phrase’, which does not further specify if the head is modified in any way. Complex nominals, however, designate a premodified and/or postmodified NP. When complex nominals are analysed, specifications of the premodification and/or postmodification patterns under study are provided.

³ The T-unit is defined as “one main clause plus the subordinate clauses attached to or embedded within it” (Hunt 1965: 49).

and Lahuerta-Martínez (2018, 2023), who included the measures noun phrases per clause and mean length of noun phrase, respectively. Only the study by Díez-Bedmar and Pérez-Paredes (2020) and Sarte and Gnevsheva (2022) paid exclusive attention to NP complexity.

Since NP complexity is typical of academic writing (Biber *et al.* 2011, 2021), text types in that writing context have been the most frequently analysed ones (e.g., argumentative texts). As a result, there is a gap in the literature regarding the study of NP complexity in other text types which are not considered academic prose, such as email writing. These studies are, in fact, necessary to analyse the effect that text type, genre, and even topic—which have been shown to affect syntactic complexity—may have on NP complexity and consequently reach a more exhaustive understanding of NP complexification in learner language (Biber and Gray 2011; Lu 2011; Polio and Park 2016; Staples *et al.* 2016; Staples and Reppen 2016; Bernardini and Gradfeldt 2019; Lan *et al.* 2019; Sarte and Gnevsheva 2022).

2.3. *Analysing lexical diversity and syntactic complexity*

The existing body of research exploring the interaction of different CAF measures is relatively limited, with many studies focusing on isolated aspects such as lexical diversity or syntactic complexity. Studies that consider different complexity types are also scarce, but some isolated hints here and there. In a 12-month longitudinal case study with an untutored L1 Turkish learner of English, Polat and Kim (2014) analysed accuracy—by means of a global measure and one which focused on the present simple tense—as well as syntactic complexity—mean length of speech unit (AS-unit)⁴; clauses per AS-unit and mean length of clauses—and lexical diversity—using the measure of lexical diversity D—in the participant’s speech in oral interviews. The results indicated that the only clear improvement was seen in the participant’s vocabulary, with increased lexical diversity. Another longitudinal study exploring lexical and syntactic complexity was conducted by Kisselev *et al.* (2022) who analysed a learner corpus of Russian with argumentative and narrative essays written by students at different proficiency levels. Lexical complexity was operationalised by means of the mean word length and the lexical frequency profile

⁴ An AS-unit stands for Analysis of Speech unit and refers to a unit of speech or writing that consists of a main clause and any subordinate clauses associated with it (Foster *et al.* 2000).

(at A1 and B2 levels). Syntactic complexity was analysed by means of measure of textual lexical diversity (MTLD) by lemma, MTLD by wordform, mean sentence length, clauses per sentence, coordinate clause ratio, subordinate clause ratio, syntactic depth ratio, relative clause ratio, infinitive clause ratio, participle clause ratio, and gerund clause ratio. The results show that nine indices —namely, mean word length, type-token ratio, percentage of high-frequency words, mean sentence length, clauses per sentence, syntactic depth, proportion of subordinate clauses, and proportion of relative clauses— showed differences in the course of the eight-week instruction programme and, for clauses per sentence, correlated with the results of the initial placement test and the final proficiency test.

Gaillat *et al.* (2022: 132) employed ‘microsystems’ —i.e., “families of competing constructions in a single paradigm”— to classify learner texts from A1 to C2 levels. Their results showed that, although the consideration of lexical (lexical variation and lexical sophistication), syntactic (syntactic complexity), and pragmatic features (cohesion) in learner writing, as well as their accuracy (considering average misspelling every 50 words) as retrieved by the software LCA (Lu 2012), TAALES (Kyle and Crossley 2015), L2SCA (Lu 2010), TAACO (Crossley *et al.* 2016) are important to predict CEFR levels and show that lexical and syntactic features play a determinative role in the prediction.

Lahmann *et al.* (2019) analysed different measures of linguistic complexity in the spontaneous oral production by German-English bilinguals living in an English-speaking country to reveal clusters of grammatical and lexical complexity measures. To do so, measures at syntactic level (sentence and clause, sub-clause, and phrase), morphological level and at lexical level (diversity and sophistication) were considered. The results reveal that the cluster for grammatical complexity measures include length measures and subordination ratios regarding grammatical complexity, without forgetting measures for sentence types and morphology. As for lexical complexity, lexical diversity, frequent lexical items, and the use of abstract words were important in the cluster for lexical complexity.

Finally, Lambert and Nakamura (2019) studied clausal (the proportion of simple utterances, compound utterances, complex utterances based on nominal subordination, adverbial subordination, and relative subordination), phrasal complexity (words per NP, modifier tokens per NP, modifier types per NP, subordinate nouns per NP) as well as the abstractness of the head nouns in the oral production by 36 L1 Japanese learners of

English at different proficiency levels, also considering the production by 18 L1 English peers. Among the results, the role played by the students' access to task-relevant lexis as a moderating variable is highlighted. Regarding the relation between phrasal complexity and lexis, the more proficient students were found to produce specific words more frequently than lower-level students, who compensated their lack of specific vocabulary with more complex NPs. In NPs whose head were the nouns 'part', 'thing', and 'place', the most frequent postmodifier was the relative clause. Lambert and Nakamura's identification of task-relevant lexis as a moderating variable aligns with some of our findings, particularly regarding how more proficient learners tend to use specific vocabulary and some NP complexification types. By incorporating these perspectives, our study aims to contribute to the ongoing discussion on language acquisition, potentially offering new angles for both theoretical exploration and practical applications in language teaching and assessment.

In conclusion, while research on the interaction of different complexity measures in learner language remains relatively scarce, the studies reported here underscore the multifaceted nature of linguistic development. Our study aims to extend these findings by examining the interplay between lexical diversity and syntactic complexity across different proficiency levels, thus providing a broader perspective. By combining diverse complexity measures, our research contributes to a more comprehensive understanding of NP learner language production.

3. METHODOLOGY

3.1. *The learner corpus*

The learner corpus used in this study is a subsection of the *FineDesc Learner Corpus*, which is being compiled within the *FineDesc Research Project*, funded by the *Spanish Ministry of Science Innovation and Universities*, with the pass-only texts by L1 Spanish candidates who have taken the high-stakes CertAcles Exam Suite at B1, B2, or C1 level in Spanish University Language Centres.⁵ In the exam, candidates are asked to write two different texts which are evaluated by two professional independent raters. Only those texts which meet the requirements of the level (B1, B2, or C1, depending on the exam

⁵ Further information on the *FineDesc Research Project* can be found at <https://web.ujaen.es/investiga/finedesc/index.php>

taken), as determined by two experienced CEFR raters who evaluate high-stakes examinations in University Language Centres in Spain, are included in the *FineDesc Learner Corpus* after they are fully anonymised and transcribed into electronic format (txt files).

The subsection considered for this study consists of 90 texts at two CEFR levels (amounting to 18,134 words), as shown in Table 1.

Level	Number of texts	Number of words	Mean and Standard Deviation
B1	44	6,795	M= 154.43 / SD= 31.36
C1	46	11339	M= 246.50 / SD= 41.04
Total	90	18,134	

Table 1: An overview of the learner corpus employed

Students at both levels were asked to react to the same prompt (reply to a friend's email), but using a different number of words, depending on the level (see Figure 1). As can be seen in the data in Table 1, students aimed at the highest number of words required per level. Candidates at B1 wrote a mean of 154.43 words (the maximum number of words required was 150) and those at C1 level produced a mean of 246.50 words (the maximum number of words required was 250). The difference in the means of words per text type proved to be statistically significant ($p = .000$; $z = -7.269$; $U = 111.500$). Normalisation of the data per 1,000 words was, therefore, calculated to compare the results across the two subcorpora.

You have just received an email from your friend Alex, who you met last year while studying abroad with an Erasmus+ grant in Switzerland. In the email, Alex mentions that s/he and their family are planning to visit your town for a few days. Alex has also invited you to give a talk to a group of his/her employees.

Write a reply in **120–150** words (B1) // **200–250** words (C1) using the following instructions:

- Recommend different family plans to do in the city.
- Thank him/her for inviting you to speak to a group of his/her employees.
- Outline the topics you will talk about.
- Suggest a Skype phone call to discuss things in more detail.

Figure 1: Prompt provided to students for the writing task

3.2. Noun lexical diversity

This study examines a total of 680 noun lexemes extracted from a corpus of B1 (44 texts; 6,795 tokens) and C1 (46 texts; 11,339 tokens) emails taken from the *FineDesc Learner*

corpus. The sample consisted of 90 emails, which were POS tagged using *Freeling* (Padró *et al.* 2010; Padró and Stanilovsky 2012). All noun lexemes were first disambiguated by means of the UKB option (Agirre *et al.* 2018) in *Freeling*, a word sense disambiguation tool, which helps identify the correct sense of a word in a given context when the word has multiple meanings, and they were later annotated using *WordNet* (Fellbaum 1998).

A total of three (direct and inherited) hypernyms were retrieved and manually annotated, creating a hierarchical list from the most specific (direct hypernyms) to up to three levels of inherited hypernyms. A direct hypernym is the immediate parent category or class of a given word (or synset) in the *WordNet* hierarchy, while inherited hypernyms include all ancestor categories or classes of a word, not just the immediate parent. The hierarchical list included up to three inherited hypernyms due to the observation that the semantic relationships provided by the first few levels of inherited hypernyms are generally sufficient to capture the essential context or distinctions relevant to the task.

In some instances, the output from *Freeling* failed to generate accurate part-of-speech (PoS) annotations. For instance, some base forms were misanalysed as nouns instead of as bare infinitives, as illustrated in Figures 2 and 3. Consequently, a manual review of the automatic annotation was necessary. Additionally, *WordNet* annotations, based on hypernyms and hyponyms, were also fine-tuned. This adjustment was essential since certain lemmas with polysemous senses —such as the distinction between ‘experience’ as an ability (background training, qualifications, etc.) and ‘experience’ as an event (something that happened on a given occasion)— had not been accurately disambiguated according to the specific context.

I	i	PRP	-	-				
,	,	Fz	-	-				
ll	ll	NN	-	-				
start	start	NN	07325190-n	0.000392536	beginning, happening, event			
with	with	IN	-	-				
a	a	DT	-	-				
review	review	NN	00879271-n	0.000294943	examination, investigation, work			
of	of	IN	-	-				
the	the	DT	-	-				
last	last	JJ	01013279-a	0.000529816				
year	year	NN	15203791-n	0.000744985	time period, fundamental quantity, measure			
achievements	achievement	NNS	00035189-n	0.00255916	action, act, event			

Figure 2: Adjustment of *Freeling*’s PoS annotation (C1 corpus, example 1)

To	to	TO	-	-			
begin	begin	VB	02608347-v	0.00116907			
,	,	Fc	-	-			
thank	thank	VB	00892315-v	0.00787451			
you	you	PRP	-	-			
so	so	RB	00117620-r	0.000869349			
much	much	JJ	01553629-a	0.00893656			
for	for	IN	-	-			
offer	offer	NN	07164546-n	0.00276078	message, communication, abstraction		
me	me	PRP	-	-			
this	this	DT	-	-			
great	great	JJ	01123879-a	0.00137994			
opportunity	opportunity	NN	14483917-n	0.00903082	possibility, being, state		
to	to	TO	-	-			
show	show	VB	00943837-v	0.00231472			
my	my	PRP\$	-	-			
ongoing	ongoing	JJ	00667822-a	0.007299			
project	project	NN	00795720-n	0.00468877	work, activity, act		
to	to	TO	-	-			
your	your	PRP\$	-	-			
employees	employee	NNS	10053808-n	0.00855556	worker, person, causal agent		

Figure 3: Adjustment of *Freeling's* PoS annotation (C1 corpus, example 2)

Subsequently, the list of nouns from the B1 corpus (a total of 304 noun lexemes) and the list from the C1 corpus (a total of 664 noun lexemes) were classified as hyponyms of the selected hypernym terms. Two databases of semantic fields (B1 and B2, respectively) were then established. The B1 database for nouns comprises 45 hypernyms, while the C1 database consists of a total of 65 hypernyms (see Appendices 1 and 2). To this respect, it is worth noting that the predictor of the semantic field of the head noun is introduced to investigate whether the context in which a noun is used influences NP complexity types. By examining the semantic fields, we aim to uncover patterns that may not be evident through a purely syntactic or lexical analysis.

Finally, the *English Vocabulary Profile* (EVP), a tool which provides detailed information about which words are typically produced by learners at each CEFR level, was consulted to verify if the noun hyponyms employed by B1 and C1 learners were classified into the B1 and C1 levels according to the tool. This study, therefore, uses the EVP to align the types of noun hyponyms used in Spanish EFL email writing with the learners' proficiency levels, providing a more precise analysis of lexical diversity.

3.3. NP complexity

In this study, NP complexity was operationalised by considering premodified NPs, postmodified NPs, and pre- and postmodified NPs (Biber *et al.* 2021: 568–642; examples are shown in Table 2—see Section 3.4 below). Since the main aim of the study is to provide an overview of NP complexity considering the effect that the head noun may have on NP complexity in email writing at different levels, no further analysis has been conducted regarding the specific linguistic structure employed by the learner to premodify, postmodify or pre- and postmodify the head of the NP.

3.4. Annotation procedure

To analyse the semantic field of the head noun in the NP and the complexity NP type employed, a two-step procedure was followed once the 2,046 NPs in the learner corpus had been identified. First, each head noun was annotated to specify the semantic field to which it belonged. To do so, the first layer of annotation was manually inserted by employing a tag which indicated one of the 36 semantic fields considered, as in examples 1 and 2:

- (1) more (communication) details about your mates (C1_8)
- (2) the (ability) methods we use to potentially increase... (C1_5)

Then, the second layer of information was added to account for the NP complexity type in each NP, as can be seen in (3) and (4) below, using the taxonomy in Table 2. As a result, the two layers of information provided both the lexical information and the syntactic complexity information necessary to conduct this study in a total of 981 NPs which showed premodification, postmodification or pre- and postmodification. The remaining 1,065 NPs in the learner corpus were determiner NPs (i.e., NPs without any complexification) and were not considered in the study.

- (3) more (communication _post) details about your mates (C1_8)
- (4) the (ability _post) methods we use to potentially increase... (C1_5)

NP complexity type	Tag	Learner example
Premodification	(prem)	(prem) Technological devices (C1_17) (prem) A brief and light <u>talk</u> (C1_18)
Postmodification	(post)	(post) <u>risk</u> of coronavirus (B1_1296) (post) <u>Advice</u> on other relevant aspects (B1_1297)
Pre- and postmodification	(prem_post)	(prem_post) New <u>things</u> about my culture (B1_1294) (prem_post) Different family <u>plans</u> to do in the city (B1_1267) (prem_post) the urban <u>bus</u> , that is quite cheap, (B1_1270) (prem_post) The worst public <u>transport</u> of Spain (B1_1295)

Table 2: The NP taxonomy employed and examples (head nouns are underlined)

To retrieve all the occurrences of each NP complexity type per semantic field, the software *Tags Retrieval* (Martínez Mimblera 2021) was employed. The information was then transferred to *SPSS* for subsequent statistical analyses. Due to the non-normal distribution of the data ($p < .05$), Mann-Whitney tests were run to single out the statistically significant differences in the use of the different NP complexity types in B1 and C1 email learner writing when using head nouns which belong to semantic fields typically attested in the B1 or C1 level.

4. RESULTS AND DISCUSSION

4.1. Noun lexical diversity

The study involved a sample of 884 noun lexemes, with a lower range of lexeme nouns in the B1 corpus (274) compared to the C1 learners' sample (610). The general results of the most prototypical hypernyms found in the B1 learner corpus are shown in Figure 4:

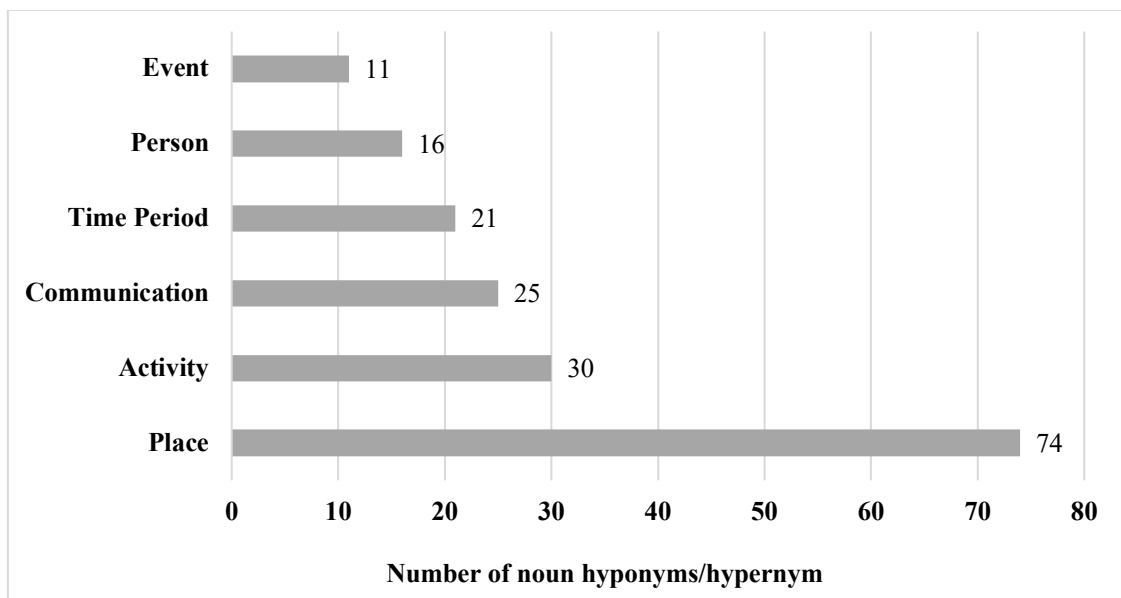


Figure 4: Most salient noun hypernyms in the B1 database

As illustrated in Figure 4 above, there are six noun hypernyms particularly used in the B1 corpus (i.e., *place*, *activity*, *communication*, *time period*, *person*, and *event*) each of them with more than ten hyponyms. These fields are closely related to the thematic situation presented in the writing task assigned to the participants, which asked them to respond to an email from a friend they had met the previous year in Switzerland. and who plans to visit Spain with their family. Additionally, the friend takes the opportunity to invite the candidate to give a talk to a group of employees in their company. The candidate is asked to address the following bullet points which trigger words from the hypernyms between brackets:

- (a) Recommend various family plans to undertake in their city (e.g., *activity*, *place*, *event*);
- (b) Express gratitude to the colleague for the invitation to speak before their workers (*communication*, *person*);
- (c) Outline the topics the candidate would like to address in the talk (*communication*, *person*);
- (d) Suggest a *Skype* meeting to discuss these matters more thoroughly (*communication*, *time period*).

A more detailed analysis of each of the six most salient hypernyms allows for the observation of the range of noun hyponyms comprising each semantic field, as shown in Figure 5 below.

Place	Activity	Communication	Time Period	Person	Event
School	Course	Question	Summer	Secretary	Landmarks
Home	Sport	Email	Week	Friend/s	Situation
University	Activity	Information	Morning	Adult/s	Visit
Camp	Travel	Advertisement	Year	Children	Matches
Hotel	Exercise	Response	Weekend	Instructors	Party/ies
Place/s	Job	Contact	Evening	Student/s	Attraction
Canteen	Exams	Answer	Hours	Players	Fair
Showers	Attention	Conclusion	Age	Roommate	Festivity
Country/ies	Help	Offer	Date	Experts	festival
Camp	Ride	Letter	Holiday/s	Parents	Event
House, houses	Plan, plans	News	Autumn	Brother	Shows
Village/s	Project	Music	Day	Cousins	
Office	Trip, trips	Advice	October	Colleague	
Area	Routes	Reply	Nights	Tourist/s	
City/ies	Scape room	Regard/	Moment	Kids	
Companies	Trekking	Ticket/s	Afternoon	Waiter	
Shops	Mountain-climbing	Story/ies	Season		
Restaurant/s	Dancing	Goodbye	Times		
Pubs	Degree	Recommendations	Month		

Figure 5: Noun hyponyms in the selected hypernyms from the B1 database

Place	Activity	Communication	Time Period	Person	Event
Cinemas	Care	Tips			
Mountain	Tourism	Card			
Monument/s	Hiking	Media			
Cathedral	Climbing	Archive			
World	Cycling	webpages			
Liverpool	Walking				
Santiago de Compostela	Shopping				
Jaén	Project				
Spain	Hugs				
Sierra Cazorla	Kisses				
Arab Baths					
Town/s					

Figure 5: Continuation

The data reveal that the most prototypical semantic fields in the B1 texts correspond to more concrete areas, consistent with the expected vocabulary associated with B1, which tends to be more concrete, practical, and focused on everyday situations. For example, the hypernym *activity* includes hyponyms such as *sport*, *hiking*, *cycling*, and *shopping*, which are typical activities one might recommend for a family visit. Similarly, the hypernym *communication* includes nouns like *email*, *call*, and *talk*, reflecting the communicative actions required by the task.

In terms of vocabulary and topics, learners at the B1 level often engage with everyday situations, personal experiences, hobbies, travel, and simple professional topics (Council of Europe 2001, 2020; North 2021). It is also noteworthy that some of the nouns used correspond to lexical items lifted directly from task instructions (e.g., *course*, *friend*, *secretary*, *school*, *accommodation*). This reliance on task-related vocabulary indicates a tendency for B1 learners to use familiar and concrete nouns that are closely tied to the given context.

By contrast, according to the EVP, the list of nouns (hyponyms) used in the B1 learner corpus predominantly corresponds to nouns associated with lower CEFR levels. As illustrated in Figure 6, the comparison between the noun database in the B1 learner corpus of email writing and the EVP shows that only 21.90 per cent of the nouns in the learner corpus are B1 nouns in the EVP, while the majority of the nouns used (62.04%) are categorised in lower levels; for instance, *school*, *summer*, *things*, *friend*, *sport*, *train*, *weather*, *job*, *food* (A1), and *office*, *story*, *team*, *moment*, *nature*, *luggage*, *information*, *exercise* (A2) in the CEFR. This indicates that B1 learners frequently use basic nouns that are well within their comfort zone and are likely to encounter in everyday interactions.

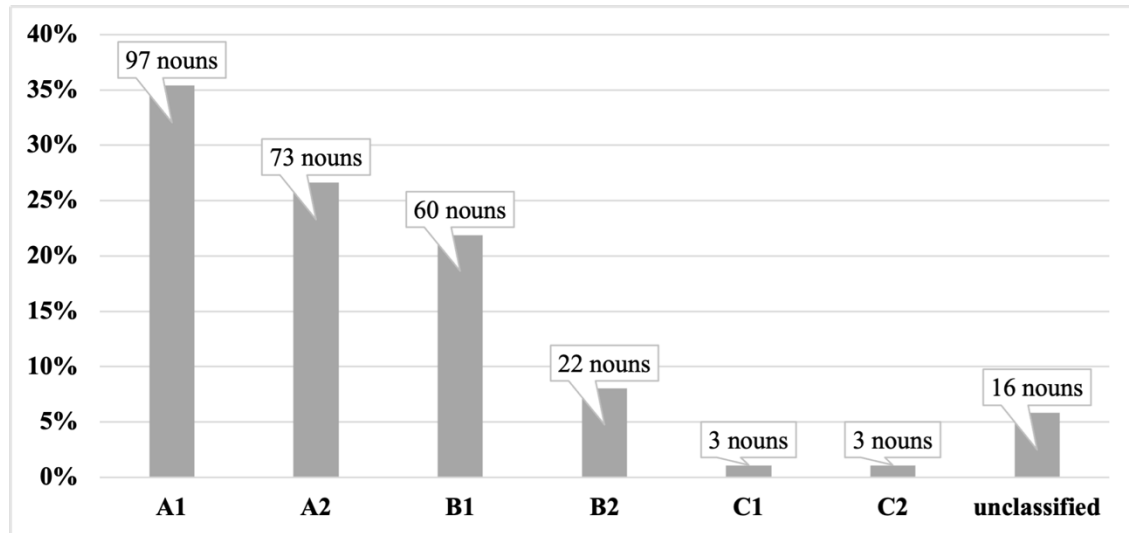


Figure 6: Comparison between the noun database in the B1 corpus and the EVP

In the C1 learner corpus, the general results regarding the most prototypical hypernyms reveal an increase in the range of nouns belonging to different semantic fields. As shown in Figure 7, the noun hyponyms associated with the hypernyms of *communication*, *activity*, *time period*, *person*, and *event* are also the most frequently used, although in different proportions. While in B1 we observed that *activity* was the most frequent, followed by *communication*, in C1, we see that the use of nouns in the *communication* field is higher than that of *activity*.

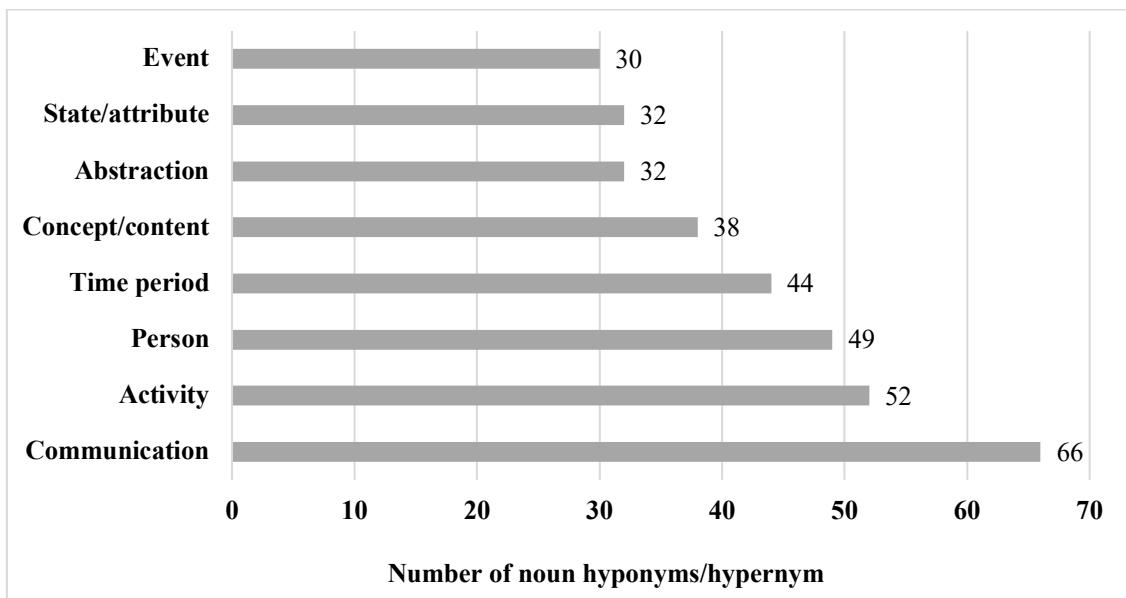


Figure 7: Most salient noun hypernyms in the C1 database

Figure 8 shows the data for the analysis of the actual hyponyms of the most salient hypernyms in the C1 corpus:

Communication	Activity	Person	Time Period	Concept/ content	Abstraction	State/ attribute	Event
Talk/s	Work/s	Employee/s	Moment	Content/s	Knowledge	Friendship	Festival
Communication	Project/s	Students	Future	Topic/s	Situation/s	Advantage/s	Event/s
(Phone) call	Activity/ies	Customers	Lifetime	Theme	Level	Pitch	Time (occasion)
Email	Preparation	Workers	Day/s	Subject/s	Way/s	Atmosphere	Impact
News	Practice	Followers	Hour/s	View/s	Time	Relationship/s	Change/s
Suggestions	Investigation/s	Mate/s	Break	Field/s	Depth	Position	Championship
Answer	Care	Doctor	School	Discipline/s	Background	Personality	Video-conference
Message	Classes	Designers	Summer	Architecture	Duration	Life	Discovery
Question/s	Education	Boy	Week/s	Technology/ies	Spirituality	Control	Performance
Speech/s	Attention	Members	Yesterday	Matter/s	Relevance	Mindfulness	Participation
Consideration	Research	Co-workers	Ages	Norms	Motivation	Quality	Accidents
Discussion	Task/s	Managers	Year/s	Science	Arrangement/s	Childhood	Championship
Encouragement	Teamwork	Husband	Month	Law	Connection	Degree	Convention
Tips	Service/s	Children	Monday/s	Goal/s	Ethics	Benefits	Implementation
Expression	Process	Kids	Morning/s	Idea/s	Relation	Attitude	Conference/s
Thanks	Career	Toddlers	Holidays	Expectation/s	Need	Doubt/s	Meeting/s
Paragraphs	Exercise/s	Producers	Date/s	Mind	Dynamics	Quantity	Demonstration
Advice	Job	Counsellors	Tomorrow	Inspiration	Protocol	Postures	Workshop/s
Songs	Hobby	Therapists	Session/s	Thought/s	Act	differences	Improvement/s
Negotiation/s	Practice	Victims	Afternoon/s	Rules	Rights	Danger	Appointment/s
Promotion	Training	Lawyer	Times	Studies	Pressure	Aspect/s	Distribution
Proposal	Tricks	Scholar	Right	Objectives	Responsibilities	Disadvantages	Spread
Conversation/s	Instruction/s	Person	Weekend/s	Aim/s	Functions	Reality	Video-conference
Voice/s	Course/s	Speaker	Decades	Issue/s	Security	Perks	Progress
Arguments	Treatment	Spectators	Today	Concept	Resources	Privilege	Steps
Information	Efforts	Experts	Fortnight	Plan/s	Rates	Wellbeing	Convention
Document	Use	Workmates	Leisure	Strategy/ies	Proactivity	Circumstances	Experience
Summary	Brainstorming	Assistants	Phases	Policy	Organisation	Reputation	Championship
Statement	Cooperation	Employer	Evening	Economy	Obligations	Account	End
Offer	Role/s	Colleagues	Minute/s	Access	Protocol	Environment	
Communications	Review	Engineers	(in) Advance	Psychology		Self-harm	
Reply	Mediation	Clients	Friday/s	Energy			
Response	Errands	Participants	Pace	Budget/s			

Figure 8: Noun hyponyms in the selected hypernyms from the C1 database

As mentioned in Section 3, C1 learners were asked to produce the same writing task as the one provided to B1 learners. Due to the task's typology and topic, the high frequency of nouns falling into the semantic fields of *communication* (e.g., *consideration*, *negotiation*, *statement*), *person* (e.g., *customer*, *followers*, *co-workers*, *assistants*, *engineers*), and *event* (e.g., *championship*, *convention*) was also noticeable in the B1 sample. However, as expected, the most notable aspect in C1 texts is the incorporation of hypernyms related to concepts that are more abstract. As shown in Figure 8, this includes categories such as *concept/content* (38 nouns), *abstraction*, and *state/attribute* (32 nouns each).

At the C1 level, learners are expected to handle more nuanced vocabulary, allowing them to engage in discussions on a wide range of topics, including those of an abstract nature. Thus, they are likely to encounter and use words that are more precise, sophisticated, and context-dependent compared to learners at the B1 level. Examples of these words are *topic*, *matter*, *expectation*, *mind*, *goal* (*concept/content*); *knowledge*,

background, relevance, arrangement (abstraction); and advantage, friendship, childhood, or mindfulness (state/attribute).

The comparison with the EVP reflects the same trend noticed in the B1 learner corpus. Only 8.03 per cent of the nouns in the C1 learner corpus are accounted for as C1 nouns in the EVP (e.g., *negotiation, innovation, relevance, outcome, motivation*) —see Figure 9— whereas the vast majority (79.84%) of nouns used are from lower levels (A1-B2), mainly B1 and B2, as shown in the following examples: *game, students, people, country, village, hotel, letter* (A1); *price, news, photograph, company, machine* (A2); *employee, knowledge, opportunity, explanation, equipment, decision* (B1); *inhabitant, need, advantage, wish, emotion, therapy, anxiety, ambition* (B2).

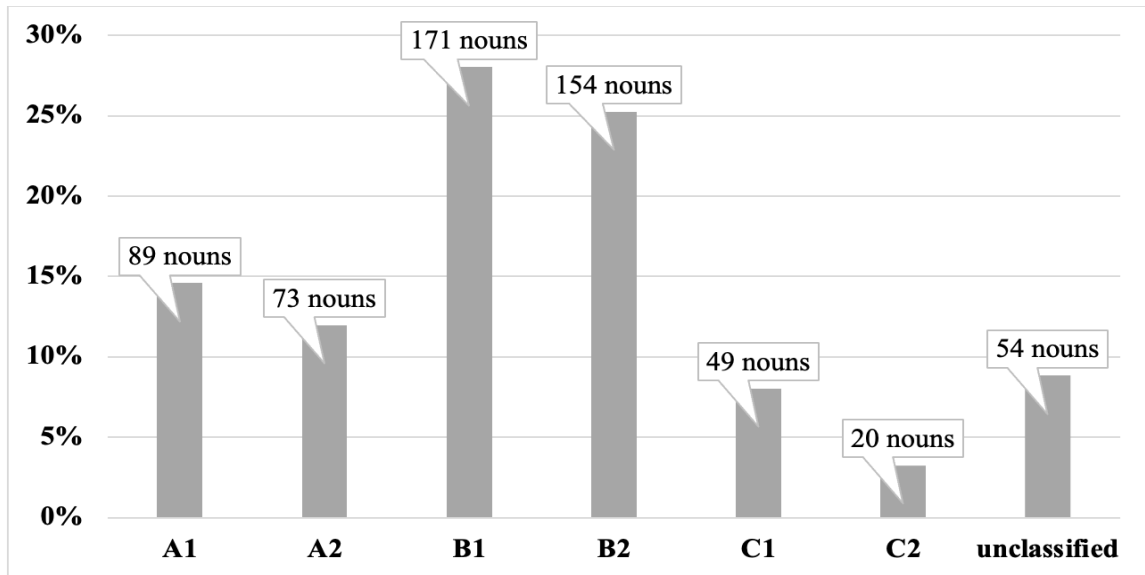


Figure 9: Comparison between the noun database in the C1 corpus and the EVP

The results confirm the initial expectation that there is a lower range of nouns in the B1 sample in comparison to the C1 learners' sample, and they reflect that the study has considered the influence of topic and text type on the use of nouns from different semantic fields. As evidenced by Du *et al.* (2022: 8), the lexical choices made by language learners may be influenced by several variables, such as text types, genres, and registers (Biber and Conrad 1999; Hyland 2008) and so may EFL learners' written production.

On the contrary, the comparison of the B1 and C1 learner subcorpora reveals unexpected results. In our sample, both at B1 and C1 levels, the percentage of nouns that, following the information in the EVP, belong to the target level is remarkably low. While the lexical choices of both B1 and C1 learners appear to be influenced by the specific tasks assigned to them, the observed distribution of nouns across CEFR levels raises

interesting questions. This finding suggests that while learners may achieve a certain proficiency level, their lexical choices are not strictly confined to that level, indicating a more fluid progression in vocabulary acquisition. Further research could explore the factors influencing this variability, such as exposure to different text types, personal interests, and the pedagogical approaches employed in teaching vocabulary.

In conclusion, our study provides valuable insights into the lexical diversity of EFL learners at different proficiency levels, highlighting the need for a nuanced understanding of vocabulary development. By examining both quantitative data and qualitative nuances, educators and researchers can better support learners in expanding their lexical repertoire in a way that aligns with their communicative needs and proficiency goals.

4.2. NP complexity

Out of the 2,046 NPs in the learner corpus, 981 NPs were manually identified as premodified NPs, postmodified NPs or pre- and postmodified NPs (the remaining NPs were determiner NPs). The analysis of NP complexity in the 981 NPs was first conducted with NPs whose head nouns belong to a semantic field typical of CEFR B1 level in email writing by B1 and C1 learners, and was then carried out the NPs whose head nouns belong to a semantic field typical of CEFR B2. Inferential statistical analyses were then run to determine differences in the use of NP complexity types considering the two variables (semantic field typical of B1 or C1 and learner level, B1 or C1), see Tables 3 and 4.

The comparison of the NP complexity types, divided into premodified NPs, postmodified NPs, and pre- and postmodified NPs, in NPs whose head nouns belong to a semantic field typical of CEFR B1 level in the production by B1 and C1 learners reveals statistically significant differences in the production —by B1 and C1 learners— regarding only six hypernyms (out of the 34 considered), namely *place*, *means of transport*, *clothing*, *time*, *utensil* and *communication* (see Table 3 for hypernyms, NP complexity types and descriptive and inferential statistics).⁶ Therefore, words related to the other semantic fields typical of B1 (see Appendix 1) are used in the different NP complexity types with a similar frequency by students at B1 and C1 levels.

⁶ The descriptive statistics reported in Tables 3 and 4 are the means (*M*), standard deviations (*SD*), medians (*Mdn*) and interquartile ranges (*IQR*), the latter provided because of the non-normal distribution of the data which required the use of non-parametric tests. Then, the results of the Mann-Whitney tests are offered by reporting the Mann-Whitney *U* value, the *z* value as well as the effect size for each comparison.

Hypernyms	Np complexity type	Statistics			
		<i>p</i> -value	B1 Data	C1 Data	Mann-Whitney test and effect size
Place	Premodified NP	<i>p</i> = .00	<i>M</i> = 9.23	<i>M</i> = .83	<i>z</i> = -5.950
			<i>SD</i> = 7.93	<i>SD</i> = 1.94	<i>U</i> = 341.000
			<i>Mdn</i> = 7.49	<i>Mdn</i> = .00	<i>r</i> = .63
			<i>IQR</i> = 14.16	<i>IQR</i> = .00	
	Postmodified NP	<i>p</i> = .00	<i>M</i> = 7.06	<i>M</i> =.37	<i>z</i> = -5.027
			<i>SD</i> = 8.53	<i>SD</i> = 1.24	<i>U</i> = 501.000
			<i>Mdn</i> = 6.00	<i>Mdn</i> = .00	<i>r</i> = .53
			<i>IQR</i> = 12.28	<i>IQR</i> = .00	
	Pre- and postmodified NP	<i>p</i> = .00	<i>M</i> = 3.88	<i>M</i> =.15	<i>z</i> = -5.194
			<i>SD</i> = 4.15	<i>SD</i> = .72	<i>U</i> = 504.000
			<i>Mdn</i> = 4.88	<i>Mdn</i> = .00	<i>r</i> = .55
			<i>IQR</i> = 6.93	<i>IQR</i> = .00	
Communication	Premodified NP	<i>p</i> =.00	<i>M</i> = 2.53	<i>M</i> = 6.62	<i>z</i> = -4.427
			<i>SD</i> = 3.52	<i>SD</i> = 4.30	<i>U</i> = 478.500
			<i>Mdn</i> = .00	<i>Mdn</i> = 7.08	<i>r</i> =.47
			<i>IQR</i> = 5.78	<i>IQR</i> = 5.93	
	Postmodified NP	<i>p</i> =.010	<i>M</i> = 2.69	<i>M</i> = 4.64	<i>z</i> = -2.579
			<i>SD</i> = 5.24	<i>SD</i> = 5.15	<i>U</i> = 723.000
			<i>Mdn</i> = .0	<i>Mdn</i> = 3.53	<i>r</i> = .27
			<i>IQR</i> = 5.81	<i>IQR</i> = 8.09	
	Pre- and postmodified NP	<i>p</i> =.022	<i>M</i> = .2850	<i>M</i> = 1.10	<i>z</i> = -2.295
			<i>SD</i> = 1.32	<i>SD</i> = 2.24	<i>U</i> = 844.000
			<i>Mdn</i> = .00	<i>Mdn</i> = .00	<i>r</i> = .24
			<i>IQR</i> = .00	<i>IQR</i> = .00	
Means of transport	Premodified NP	<i>p</i> =.00	<i>M</i> =3.26	<i>M</i> =.18	<i>z</i> = -4.304
			<i>SD</i> = 4.38	<i>SD</i> = .84	<i>U</i> = 943.000
			<i>Mdn</i> = .00	<i>Mdn</i> = .00	<i>r</i> = .45
			<i>IQR</i> = 6.78	<i>IQR</i> = .00	
Clothing	Premodified NP	<i>p</i> =.019	<i>M</i> =.78	<i>M</i> = .00	<i>z</i> = -2.338
			<i>SD</i> = 2.21	<i>SD</i> = .00	<i>U</i> = 897.000
			<i>Mdn</i> = .00	<i>Mdn</i> = .00	<i>r</i> = .25
			<i>IQR</i> = .00	<i>IQR</i> = .00	
Time	Premodified NP	<i>p</i> =.001	<i>M</i> =1.22	<i>M</i> =3.03	<i>z</i> = -3.414
			<i>SD</i> = 3.05	<i>SD</i> = 2.87	<i>U</i> = 643.500
			<i>Mdn</i> = .00	<i>Mdn</i> = 3.72	<i>r</i> = .36
			<i>IQR</i> = .00	<i>IQR</i> = 4.74	
Utensil	Premodified NP	<i>p</i> =.047	<i>M</i> =.00	<i>M</i> =.37	<i>z</i> = -1.989
			<i>SD</i> = .00	<i>SD</i> = 1.22	<i>U</i> = 924.000
			<i>Mdn</i> = .00	<i>Mdn</i> = .00	<i>r</i> = .21
			<i>IQR</i> = .00	<i>IQR</i> = .00	

Table 3: Hypernyms related to B1 which show statistically significant differences in the use of some NP complexity type by B1 and C1 learners

Second, the variety of NP complexity types which shows statistically significant differences in learner writing at B1 and C1 level per semantic field varies. As seen in Table 3, the head nouns corresponding to the hypernyms *place* and *communication* are found in premodified NPs, postmodified NPs and pre- and postmodified NPs which are more frequently employed either by C1 learners (in the case of *communication* head nouns) or B1 learners (with *place* head nouns), in the latter case with large effect sizes in

all complexification types. The data point to B1 learners' use of the three NP complexity types to describe places —see (5) and (6)— whereas the case is so for C1 learners with the head nouns related to communication, as illustrated in (7) and (8).

(5) taking one of the school (places_prem) courses (B1_1182)

(6) the (places_post) city where you and your family could enjoy... (B1_1247)

(7) give a (communication_post) talk about “How to improve your teaching” (C1_200002)

(8) the main (communication_prem_post) outlines of my speech (C1_200017)

However, the head nouns which belong to the other four hypernyms, namely *utensil*, *time*, *means of transport* and *clothing*, are found to differ in their frequency of use in premodified NPs only: learners at B1 level describe head nouns related to *clothing* and *means of transport* by means of premodification more than their C1 counterparts, whereas C1 learners do so with head nouns related to *utensil* and *time*, more abstract hypernyms. The data in Table 3 also reveal that C1 students do not further describe *clothing* by means of premodification and B1 students do not do so with *utensils*.

Since the students at both levels were provided with the same writing task, no text-type effect or topic effect (cf. Biber and Gray 2011; Lu 2011; Polio and Park 2016; Staples *et al.* 2016; Staples and Reppen 2016; Bernardini and Gradfeldt 2019; Lan *et al.* 2019; Sarte and Gnevsheva 2022) may explain the differences in use of the NP complexity types when using head nouns in different semantic fields related to B1. It may be the case, however, that students at the lower level are more familiar with B1 words which have a more factual/concrete meaning, such as those describing *places*, *clothing* and *means of transport*, which favours their use and the complexification of the NPs in which they are head nouns, as illustrated in (9) and (10).

(9) wear winter (clothing_prem) clothes (B1_1280)

(10) ... The public (means_of_transport_prem) transport is easy to use (B1_1279)

Our data reveal that the C1 learners, however, complexify more frequently the NPs whose head nouns express more abstract or specific words, i.e., those expressing *time*, *communication*, and *utensils*. They mainly do so by means of premodification—as shown in (11) and (12)— even though their higher language communicative competence would enable C1 students to employ any other complexification type (i.e., postmodification and/or pre- and postmodification), as they do with *communication* head nouns.

(11) because of modern (utensil_prem) instruments (C1_2000059)

(12) evening (time_prem) time (C1_200060)

This tendency suggests that complexification is lexically triggered, as the semantic properties of the head nouns have an effect on NP syntactic complexity. The use of different head nouns and complexification patterns at B1 and C1 levels characterise NP learner use at these two CEFR levels.

The data in Table 4 below show that nouns in nine semantic fields prototypical of CEFR C1 level were found in NP complexity types which show statistically significant differences in their use by B1 and C1 learners. In all cases, C1 learners employed the NP complexity types with those head nouns more frequently than B1 learners. These semantic fields are *concept/content*, *abstraction*, *entity*, *ability*, *difficulty*, *state/attribute*, *possibility*, *social control*, and *commerce/exchange*.

Hypernyms	NP complexity type	Statistics			
		<i>p</i> -value	B1 Data	C1 Data	Mann-Whitney test and effect size
Statue/attribute	Premodified NP	<i>p</i> = .00	<i>M</i> = .53	<i>M</i> = 3.01	<i>z</i> = -3.732
			<i>SD</i> = 1.71	<i>SD</i> = 4.21	<i>U</i> = 647.000
			<i>Mdn</i> = .00	<i>Mdn</i> = .00	<i>r</i> = .39
			<i>IQR</i> = .00	<i>IQR</i> =4.51	
	Postmodified NP	<i>p</i> =.00	<i>M</i> = .00	<i>M</i> = 1.32	<i>z</i> = -3.607
			<i>SD</i> = .00	<i>SD</i> = 2.44	<i>U</i> = 748.000
			<i>Mdn</i> = .00	<i>Mdn</i> = .00	<i>r</i> = .38
			<i>IQR</i> = .00	<i>IQR</i> =3.22	
	Pre- and postmodified NP	<i>p</i> =.007	<i>M</i> = .00	<i>M</i> = .7151	<i>z</i> = -2.677
			<i>SD</i> = .00	<i>SD</i> = 1.76	<i>U</i> = 858.000
			<i>Mdn</i> = .00	<i>Mdn</i> = .00	<i>r</i> = .28
			<i>IQR</i> = .00	<i>IQR</i> = .00	
Concept/content	Premodified NP	<i>p</i> =.00	<i>M</i> = .65	<i>M</i> = 3.39	<i>z</i> = -4.975
			<i>SD</i> = 2.09	<i>SD</i> = 2.99	<i>U</i> = 475.000
			<i>Mdn</i> = .00	<i>Mdn</i> = 3.60	<i>r</i> = .52
			<i>IQR</i> = .00	<i>IQR</i> =4.70	
	Postmodified NP	<i>p</i> =.00	<i>M</i> = .46	<i>M</i> = 5.45	<i>z</i> = -5.514
			<i>SD</i> = 1.73	<i>SD</i> = 6.21	<i>U</i> = 422.000
			<i>Mdn</i> = .00	<i>Mdn</i> = 3.77	<i>r</i> = .58
			<i>IQR</i> = .00	<i>IQR</i> = 8.68	
	Pre- and postmodified NP	<i>p</i> =.40	<i>M</i> = .88	<i>M</i> = 1.40	<i>z</i> = -2.058
			<i>SD</i> = 2.68	<i>SD</i> = 2.20	<i>U</i> = 826.500
			<i>Mdn</i> = .00	<i>Mdn</i> = .00	<i>r</i> = .22
			<i>IQR</i> = .00	<i>IQR</i> = 3.53	

Table 4: Hypernyms related to C1 which show statistically significant differences in the use of some NP complexity types by B1 and C1 learners

Hypernyms	NP complexity type	Statistics			
		<i>p</i> -value	B1 Data	C1 Data	Mann-Whitney test and effect size
Ability	Premodified NP	<i>p</i> =.002	<i>M</i> =.00 <i>SD</i> = .00 <i>Mdn</i> = .00 <i>IQR</i> = .00	<i>M</i> =.93 <i>SD</i> = 2.04 <i>Mdn</i> = .00 <i>IQR</i> = .00	<i>z</i> = -3.070 <i>U</i> = 814.000 <i>r</i> = .32
	Pre- and postmodified NP	<i>p</i> =.002	<i>M</i> =.00 <i>SD</i> = .00 <i>Mdn</i> = .00 <i>IQR</i> = .00	<i>M</i> =.76 <i>SD</i> = 1.58 <i>Mdn</i> = .00 <i>IQR</i> = .00	<i>z</i> = -3.070 <i>U</i> = 814.000 <i>r</i> = .32
Social control	Premodified NP	<i>p</i> =.047	<i>M</i> = .00 <i>SD</i> = .00 <i>Mdn</i> = .00 <i>IQR</i> = .00	<i>M</i> =.45 <i>SD</i> = 1.53 <i>Mdn</i> = .00 <i>IQR</i> = .00	<i>z</i> =-1.989 <i>U</i> = 924.000 <i>r</i> = .21
	Premodified NP	<i>p</i> =.047	<i>M</i> = .00 <i>SD</i> = 0 <i>Mdn</i> = 0 <i>IQR</i> = 0	<i>M</i> =.45 <i>SD</i> = 1.592 <i>Mdn</i> = .00 <i>IQR</i> = 0	<i>z</i> = -1.989 <i>U</i> = 924.000 <i>r</i> = .21
Possibility	Postmodified NP	<i>p</i> =.023	<i>M</i> =.42 <i>SD</i> = 2.00 <i>Mdn</i> = .00 <i>IQR</i> = .00	<i>M</i> = 1.05 <i>SD</i> = 2.08 <i>Mdn</i> = .00 <i>IQR</i> = .00	<i>z</i> = -2.268 <i>U</i> = 846.000 <i>r</i> = .24
	Postmodified NP	<i>p</i> =.028	<i>M</i> = 1.69 <i>SD</i> = 3.473 <i>Mdn</i> = 0 <i>IQR</i> = 0	<i>M</i> = 2.83 <i>SD</i> = 3.460 <i>Mdn</i> = 3.20 <i>IQR</i> = 4	<i>z</i> = -2.196 <i>U</i> = 773.000 <i>r</i> = .23
Difficulty	Postmodified NP	<i>p</i> =.047	<i>M</i> =.00 <i>SD</i> = .00 <i>Mdn</i> = .00 <i>IQR</i> = .00	<i>M</i> =.37 <i>SD</i> = 1.22 <i>Mdn</i> = .0000 <i>IQR</i> = .00	<i>z</i> = -1.989 <i>U</i> = 924.000 <i>r</i> = .20
	Postmodified NP	<i>p</i> =.005	<i>M</i> =.77 <i>SD</i> = 2.20 <i>Mdn</i> = .00 <i>IQR</i> = .00	<i>M</i> = 1.99 <i>SD</i> = 2.57 <i>Mdn</i> = .00 <i>IQR</i> = 4.42	<i>z</i> = -2.806 <i>U</i> = 741.5000 <i>r</i> = .30

Table 4: Continuation

As was the case with the nouns in semantic fields related to B1, differences are found regarding the variety of NP complexity types in which head nouns classified into semantic fields related to C1 are statistically more frequently used by one learner group (C1 in all cases in which words related to C1 are considered). The hyponyms of the hypernyms *concept/content* and *state/attribute* are head nouns in the three possible NP complexity types (premodified NPs, postmodified NPs, and pre- and postmodified NPs), and the words related to *ability* are used as head nouns in premodified NPs and pre- and postmodified NPs. The words related to the other six semantic fields are employed as head nouns in only one NP complexity type whose use is higher in C1 production: *social control* and *commerce/exchange* are present in premodified NPs, whereas *possibility*, *abstraction*, *difficulty* and *entity* are in postmodified NPs. On some occasions, these

differences are due to the B1 students' decision not to employ NP complexity types with head nouns related to specific semantic fields typical of C1 level, contrarily to their complexification of head nouns related to the semantic fields in B1 (see Table 3). Table 4 shows that B1 learners do not postmodify or pre- and postmodify NPs in which the head refers to *states/attributes*, but rather premodify head nouns in that semantic field. In the case of the head nouns regarding *ability*, B1 students do not premodify or pre- and postmodify such heads. Premodified NPs are not found in the production by B1 students with noun-heads referring to *social control* and *commerce/exchange* and postmodified NPs are not employed by this learner group when expressing *difficulty*. These findings highlight that C1 students are ready to premodify, postmodify and pre- and postmodify head nouns in these semantic fields, whereas B1 students show some restrictions in their way to refer to *states/attributes*, *ability*, *social control*, *commerce/exchange*, and *difficulty*.

The data in Table 4 also reveal that differences in complexification of NPs whose head nouns are related to C1 level mainly involve postmodification,⁷ which C1 students produce more, either in postmodified NPs or in pre- and postmodified NPs. In line with the results by Sarte and Gnevsheva (2022) on the topic effect on NP complexity in which they concluded that more cognitively demanding topics favoured postmodification over premodification, the higher degree of abstractness of the nouns in semantic fields related to C1 may have triggered NP complexification to describe the referent better (cf. Lambert and Nakamura 2019) and have recurred to postmodification to do so.

In summary, the analysis of NP complexity and the head nouns reveals that B1 students complexify more concrete head nouns, whereas C1 learners complexify more abstract head nouns. When head nouns are concrete, both learner groups mainly use premodification. However, when the head nouns are abstract, B1 learners may not complexify those NPs, whereas C1 students do. With abstract head nouns, both learner groups mainly employ NP complexity types which involve postmodification, even though B1 learners do so much less frequently than their C1 counterparts.

The more frequent use of complexification types which involve postmodification with abstract head nouns is found to be in line with the findings by Sarte and Gnevsheva (2022) and Lambert and Nakamura (2019). It can be then claimed that there is a relation

⁷ In fact, only the hypernyms *social control* and *commerce/exchange* are not present in statistically significant NP complexity types which involve postmodification.

between the type of head nouns complexified and the complexification pattern employed. Furthermore, the data in our study also demonstrated that the students' communicative language level also plays a fundamental role in both the selection of the words and the complexification patterns employed.

Further studies are needed to explore the role of these two CAF measures (lexical diversity and NP complexity) and communicative language level with learner corpora compiled with students' from different L1s, proficiency levels, and this and other text types.

5. CONCLUSIONS

This study provides valuable information regarding the interaction of two complexity measures —namely, noun lexical diversity and NP syntactic complexity— in L1 Spanish EFL NP production at two CEFR levels. In doing so, this paper contributes to the still limited existing literature, which considers different lexical and syntactic measures to reach a more comprehensive understanding of NP learner production. Our data also offer insights from the analysis of an underexplored text type-email writing. This contribution fosters the analysis of non-academic text types that are crucial for meeting learners' communicative needs and improving their performance in high-stakes language exams. While the study of the NP in email writing might present certain challenges for studying nominal diversity and syntactic complexity compared to more formal or extended text types, it also sheds light on language use in everyday written interactions. We believe that further research on registers which are considered non-academic is necessary so that we can reach a comprehensive understanding of NP use in learner English.

This investigation has answered the two research questions posed. Firstly, regarding RQ1, our analysis reveals that both B1 and C1 learners demonstrate limited use of nouns appropriate for their respective CEFR levels as per the EVP database. Secondly, concerning RQ2, our findings indicate that NP complexity is indeed influenced by the semantic fields of the head noun: B1 learners tend to exhibit more frequently NP complexification with concrete and factual nouns, while C1 learners show increased NP complexification, primarily through postmodification, with more abstract nouns. Our results show that the student's selection of the head noun (concrete vs. abstract) interacts

with the NP complexity type employed (pre- or postmodification). This interaction is found to be different at the two CEFR levels analysed.

Due to the different lexical and syntactic measures employed in the previous literature, the comparison of our results with those in the existing literature is limited. However, our results are in line with previous research which points to the relation between the semantic nature of the head noun and the NP complexity type employed (Lambert and Nakamura 2019; Sarte and Gnevsheva 2022). By analysing two CEFR levels, our study further explores how this lexical-syntactic relationship changes at B1 and C1 levels.

This paper offers results which may inform language teaching and learning as well as language assessment. First, language instruction and teaching materials may focus on the acquisition and production of vocabulary typical of each level and the different ways to complexify the NP. Language assessment can also benefit from the findings in this paper, as it provides a comprehensive understanding of the relevance of both noun lexical diversity and NP complexity in characterizing the writing proficiency level of L1 Spanish learners of English.

In summary, the implications of this study underline the importance of integrating activities and instructional approaches that promote and assess lexical diversity and NP syntactic complexity from a lexicogrammatical perspective to support the development of proficiency in EFL writing.

REFERENCES

- Agirre, Eneko, Oier López de Lacalle and Aitor Soroa. 2018. The risk of sub-optimal use of Open-source NLP Software: UKB is inadvertently state-of-the-art in knowledge-based WSD. *NLP-OSS workshop at ACL* (arXiv:1805.04277). <https://doi.org/10.18653/v1/W18-2505>
- Ai, Haiyang and Xiaofei Lu. 2013. A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In Ana Díaz-Negrillo, Nicolas Ballier and Paul Thompson eds. *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam: John Benjamins, 249–264.
- Allaw, Elissa. 2021. A learner corpus analysis: Effects of task complexity, task type, and L1 and L2 similarity on propositional and linguistic complexity. *International Review of Applied Linguistics in Language Teaching* 59/4: 569–604.
- Ansarifar, Ahmad, Hesamoddin Shahriari and Reza Pishghadam. 2018. Phrasal complexity in academic writing: A comparison of abstracts written by graduate students and expert writers in applied linguistics. *Journal of English for Academic Purposes* 31: 58–71.

- Bernardini, Petra and Jonas Granfeldt. 2019. On cross-linguistic variation and measures of linguistic complexity in learner texts: Italian, French and English. *International Journal of Applied Linguistics* 29/2: 211–232.
- Biber, Douglas and Susan Conrad. 1999. Lexical bundles in conversation and academic prose. *Language and Computers* 26: 181–190.
- Biber, Douglas and Bethany Gray. 2011. Grammatical change in the noun phrase: The influence of written language use. *English Language and Linguistics* 15/2: 223–250.
- Biber, Douglas, Bethany Gray and Kornwipa Poonpon. 2011. Characteristics of conversation to measure complexity in L2 writing development. *TESOL Quarterly* 45/1: 5–35.
- Biber, Douglas, Stig Johansson, Geoffrey N. Leech, Susan Conrad and Edward Finegan. 2021. *Grammar of Spoken and Written English*. Amsterdam: John Benjamins.
- Bulté, Bram and Alex Housen. 2012. Defining and operationalising L2 complexity. In Alex Housen, Folkert Kuiken and Ineke Vedder eds. *Dimensions of L2 Performance and Proficiency: Investigating Complexity, Accuracy and Fluency in SLA*. Amsterdam: John Benjamins, 21–46.
- Bulté, Bram and Alex Housen. 2014. Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing* 26: 42–65.
- Casal, J. Elliott and Joseph J. Lee. 2019. Syntactic complexity and writing quality in assessed first-year L2 writing. *Journal of Second Language Writing* 44: 51–62.
- Clavel-Arroitia, Begoña and Barry Pennock-Speck. 2021. Analysing lexical density, diversity, and sophistication in written and spoken telecollaborative exchanges. *Computer Assisted Language Learning Electronic Journal* 22/3: 230–250.
- Council of Europe. 2001. *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
- Council of Europe. 2020. *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume*. Strasbourg: Council of Europe Publishing.
- Crossley, Scott A., Kristopher Kyle and Danielle S. McNamara. 2016. The tool for the Automatic Analysis of Text Cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior research methods* 48: 1227–1237.
- Crossley, Scott A. and Danielle S. McNamara. 2014. Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing* 26: 66–79.
- Crossley, Scott A., Tom Salsbury, Danielle S. McNamara and Scott Jarvis. 2011. Predicting lexical proficiency in language learner texts using computational indices. *Language Testing* 28/4: 561–580.
- Díez-Bedmar, María Belén. 2015. Article use and criterial features in Spanish EFL writing. In Marcus Callies and Sandra Götz eds. *Learner Corpora in Language Testing and Assessment*. Amsterdam: John Benjamins, 163–190.
- Díez-Bedmar, María Belén and Pascual Pérez-Paredes. 2020. Noun phrase complexity in young Spanish EFL learners' writing: Complementing syntactic complexity indices with corpus analyses. *International Journal of Corpus Linguistics* 25/1: 1–33.
- Du, Xiangtao, Muhammad Afzaal and Hind Al Fadda. 2022. Collocation Use in EFL Learners' Writing Across Multiple Language Proficiencies: A Corpus-Driven Study. *Frontiers in Psychology* 13: 752134. <https://doi.org/10.3389/fpsyg.2022.752134>

- 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: 59–84.
- Engber, Cheryl A. 1995. The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing* 4/2: 139–155.
- Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. Cambridge: The MIT Press.
- Foster, Pauline, and Peter Skehan. 1996. The influence of planning and task type on second language performance. *Studies in Second Language Acquisition* 18: 299–324.
- Foster, Pauline, Alan Tonkyn and Gill Wigglesworth. 2000. Measuring spoken language: A unit for all reasons. *Applied Linguistics* 21/3 : 354–375.
- Gaillat, Thomas, Andre Simpkin, Nicolas Ballier, Bernardo Stearns, Annanda Sousa, Manon Bouyé and Manel Zarrouk. 2022. Predicting CEFR levels in learners of English: The use of microsystem criteria features in a machine learning approach. *ReCALL* 34/2: 130–146.
- Gregori-Signes, Carmen and Begoña Clavel-Arroitia. 2015. Analysing lexical density and lexical diversity in university students' written discourse. *Procedia-Social and Behavioral Sciences* 198: 546–556.
- Housen, Alex, Els Schoonjans, Sonja Janssens, Aurélie Welcomme, Ellen Schoonheere and Michel Pierrard. 2011. Conceptualizing and measuring the impact of contextual factors in instructed SLA — the role of language prominence. *International Review of Applied Linguistics in Language Teaching* 49/2: 83–112.
- Howarth, Peter. 1998. The phraseology of learners' academic writing. In Anthony P. Cowie ed. *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press, 161–186.
- Hunt, Kellogg W. 1965. *Grammatical Structures Written at Three Grade Levels*. National Council of Teachers of English Research Report No. 3. Champaign: Office of Education CTE Champaign.
- Hyland, Ken. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27: 4–21.
- Ionin, Tania and María Belén Díez-Bedmar. 2021. Article use in Russian and Spanish learner writing at CEFR B1 and B2 Levels: Effects of proficiency, native language, and specificity. In Bert Le Bruyn and Magali Paquot eds. *Learner Corpus Research Meets Second Language Acquisition*. Cambridge: Cambridge University Press, 10–38.
- Khushik, Ghulam Abbas and Ari Huhta. 2020. Investigating syntactic complexity in EFL learners' writing across Common European Framework of Reference levels A1, A2 and B1. *Applied Linguistics* 41/4: 506–532.
- Kim, Jungyeon. 2021. Measuring NP complexity in Korean EFL writing across CEFR levels A2, B1 and B2. *Korean Journal of English Language and Linguistics* 21: 341–358.
- Kisselev, Olesya, Rossina Soyan, Dmitrii Pastushenkov and Jason Merrill. 2022. Measuring writing development and proficiency gains using indices of lexical and syntactic complexity: Evidence from longitudinal Russian learner corpus data. *The Modern Language Journal* 106/4: 798–817.
- Kuiken, Folkert and Ineke Vedder. 2019. Syntactic complexity across proficiency and languages: L2 and L1 writing in Dutch, Italian and Spanish. *International Journal of Applied Linguistics* 29/2: 192–210.

- Kuiken, Folkert, Ineke Vedder and Roger Gilabert. 2010. Communicative adequacy and linguistic complexity in L2 writing. In Inge Bartning, Maisa Martin and Ineke Vedder eds. *Communicative, Proficiency and Linguistic Development: Intersections between SLA and Language Testing Research*. Stockholm: European Second Language Association, 81–100.
- Kyle, Kristopher. 2016. *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-based Indices of Syntactic Sophistication*. Atlanta, GA: Georgia State University dissertation.
- Kyle, Kristopher and Scott A. Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly* 49/4: 757–786.
- Kyle, Kristopher and Scott A. Crossley. 2018. Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal* 102/2: 339–349.
- Lahmann, Cornelia, Rasmus Steinkrauss and Monika S. Schmid. 2019. Measuring linguistic complexity in long-term L2 speakers of English and L1 attriters of German. *International Journal of Applied Linguistics* 29: 173–191.
- Lahuerta Martínez, Ana Cristina. 2018. Analysis of syntactic complexity in secondary education EFL writers at different proficiency levels. *Assessing Writing* 35: 1–11.
- Lahuerta Martínez, Ana Cristina. 2023. Analysis of changes in L2 writing over the time of a short-term academic English programme. *Porta Linguarum* 39: 111–127.
- Lahuerta Martínez, Ana Cristina. 2024. The role of syntactic and lexical complexity in undergraduate writing quality. *Ibérica* 47: 251–274.
- Lambert, Craig and Sachiko Nakamura. 2019. Proficiency-related variation in syntactic complexity: A study of English L1 and L2 oral descriptive discourse. *International Journal of Applied Linguistics* 29/2: 248–264.
- Lan, Ge, Kyle Lucas and Yachao Sun. 2019. Does L2 writing proficiency influence noun phrase complexity? A case analysis of argumentative essay written by Chinese students in a first-year composition course. *System* 85: 1–13.
- Lan, Ge, Qiusi Zhang, Kyle Lucas, Yachao Sun and Jie Gao. 2022. A corpus-based investigation on noun phrase complexity in L1 and L2 English writing. *English for Specific Purposes* 67: 4–17.
- Liu, Liming and Lan Li. 2016. Noun phrase complexity in EFL academic writing: A corpus-based study of postgraduate academic writing. *The Journal of Asia TEFL* 13/1: 48–65.
- Lu, Xiaofei. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15: 474–496.
- Lu, Xiaofei. 2011. 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. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal* 96/2: 190–208.
- Lu, Xiaofei and Jifeng Wu. 2022. Noun-phrase complexity measures in Chinese and their relationship to L2 Chinese writing quality: A comparison with topic-comment-unit-based measures. *The Modern Language Journal* 106/1: 267–283.
- Martínez Mimbrera, Francisco Javier. 2021. *Tags Retrieval* [Computer software]. Jaén: Universidad de Jaén.
- Mazgutova, Diana and Judit Kormos. 2015. Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing* 29: 3–15.

- Nation, Ian S. P. 2001. *Learning Vocabulary in nother Language*. Cambridge: Cambridge University Press.
- Norris, John M. and Lourdes Ortega. 2009. Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics* 30: 555–578.
- North, Brian. 2021. The CEFR Companion Volume—What’s new and what might it imply for teaching/learning and for assessment? *CEFR Journal: Research and Practice* 4: 5–24.
- Ong, Justina and Lawrence Jun Zhang. 2010. Effects of task complexity on the fluency and lexical complexity in EFL students’ argumentative writing. *Journal of Second Language Writing* 19/4: 218–233.
- Ortega, Lourdes. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics* 4/4: 492–518.
- Padró, Lluís, Samuel Reese, Eneko Agirre and Aitor Soroa. 2010. Semantic services in FreeLing 2.1: WordNet and UKB. In Pushpak Bhattacharyya, Christiane D. Fellbaum and Piek Vossen eds. *Principles, Construction, and Application of Multilingual Wordnets*. Mumbai: Narosa Publishing House, 99–105.
- Padró, Lluís and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerk, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk and Stelios Peiperidis eds. *Proceedings of the 8th the Language Resources and Evaluation Conference*. Istanbul: European Language Research Asociation, 2473–2479.
- Paquot, Magali. 2019. The phraseological dimension in interlanguage complexity research. *Second Language Research* 35/1: 121–145.
- Parkinson, Jean and Jill Musgrave. 2014. Development of noun phrase complexity in the writing of English for Academic Purposes students. *Journal of English for Academic Purposes* 14: 48–59.
- Peters, Elke. 2016. The learning burden of collocations: The role of interlexical and intralexical factors. *Language Teaching Research* 20: 113–138.
- Polat, Brittany and Youjin Kim. 2014. Dynamics of complexity and accuracy: A longitudinal case study of advanced untutored development. *Applied Linguistics* 35/2: 184–207.
- Polio, Charlene and Ji-Hyun Park. 2016. Language development in second language writing. In Rosa María Manchón and Paul Kei Matsuda eds. *Handbook of Second Language Writing*. New York: Routledge, 287–306.
- Qin, Wenjuan and Paola Uccelli. 2020. Beyond linguistic complexity: Assessing register flexibility in EFL writing across contexts. *Assessing Writing* 45: 100465. <https://doi.org/10.1016/j.asw.2020.100465>
- Ravid, Dorit and Ruth A. Berman. 2010. Developing noun phrase complexity at school age: A text-embedded cross-linguistic analysis. *First Language* 30/1: 3–26.
- Sarte, Kayla Marie and Ksenia Gnevshva. 2022. Noun phrasal complexity in ESL written essays under a constructed-response task: Examining proficiency and topic effects. *Assessing Writing* 51: 100595. <https://doi.org/10.1016/j.asw.2021.100595>
- Šišková, Zdislava. 2012. Lexical richness in EFL students’ narratives. *Language Studies Working Papers* 4: 26–36.
- Skehan, Peter. 1989. *Individual differences in Second Language Learning*. London: Hoder Arnold.
- Staples, Shelly and Randi Reppen. 2016. Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings. *Journal of Second Language Writing* 32: 17–35.

- Staples, Shelley, Jesse Egbert, Douglas Biber and Bethany Gray. 2016. Academic writing development at university level: Phrasal and clausal complexity across level of study, discipline and genre. *Written Communication* 33/2: 149–183.
- Su, Yanfang, Kanglong Liu, Fengkai Liu, John Lee and Tan Jin. 2023. Lexical complexity in exemplar EFL texts: Towards text adaptation for 12 grades of basic English curriculum in China. *International Review of Applied Linguistics in Language Teaching* 62/1: 137–164.
- Treffers-Daller, Jeanine, Patrick Parslow and Shirley Williams. 2018. Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics* 39/3: 302–327.
- Vedder, Ineke and Veronica Benigno. 2016. Lexical richness and collocational competence in second-language writing. *International Review of Applied Linguistics in Language Teaching* 54/1: 23–42.
- Vidal, Karina and Scott Jarvis. 2020. Effects of English-medium instruction on Spanish students' proficiency and lexical diversity in English. *Language Teaching Research* 24/5: 568–587.
- Vyatkina, Nina. 2013. Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus. *The Modern Language Journal* 97/S1: 11–30.
- Wolfe-Quintero, Kathryn Elizabeth, Shunji Inagaki and Hae-Young Kim. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy and Complexity*. Honolulu: University of Hawai'i at Manoa.
- Xu, Lirong. 2019. Noun phrase complexity in integrated writing produced by advanced Chinese EFL learners. *Papers in Language Testing and Assessment* 8/1: 31–51.
- Yoon, Hyung-Jo. 2017. Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System* 66: 130–141.
- Zhang, Xiaopeng and Xiaofei Lu. 2022. Revisiting the predictive power of traditional vs. fine-grained syntactic complexity indices for L2 writing quality: The case of two genres. *Assessing Writing* 51: 100597. <https://doi.org/10.1016/j.asw.2021.100597>

Corresponding author

Natalia Judith Laso Martín

University of Barcelona

Department of Modern Languages and Literatures and English Studies

Gran Via de les Corts Catalanes, 585

08007 Barcelona

Spain

E-mail: njlaso@ub.edu

received: May 2024
accepted: October 2024

APPENDICES

Appendix 1: B1 Hypernym database

	List of Hypernyms	Number of hyponyms		List of Hypernyms	Number of hyponyms
1	Place	74	24	State/attribute	3
2	Activity	30	25	Title	3
3	Communication	25	26	Category	2
4	Time period	21	27	Condition	2
5	Person	16	28	Language unit	2
6	Event	11	29	Meal	2
7	Means of transport	9	30	Navigational system	2
8	Abstraction	8	31	System of measurement	2
9	Entity	8	32	Ability	1
10	Clothing	7	33	Charge	1
11	Concept/content	7	34	Container	1
12	Feeling	7	35	Creation	1
13	Food/drink	7	36	Decision-making process	1
14	Group	6	37	Game equipment	1
15	Proper names	6	38	Housing	1
16	Body part	5	39	Imagination	1
17	Animal	4	40	Incentive	1
18	Creation	4	41	Knowledge	1
19	Value	4	42	Precious stone	1
20	Difficulty	3	43	Representation	1
21	Environment	3	44	Software	1
22	Game	3	45	Software	1
23	Possibility	3	46	Trademark	1
Total number of noun lexemes 304					

Appendix 2: C1 Hypernym database

	List of Hypernyms	Number of hyponyms		List of Hypernyms	Number of hyponyms
1	Communication	66	34	Process	4
2	Activity	52	35	(Financial) Gain	3
3	Person	49	36	Accomplishment	3
4	Time period	44	37	Beginning	3
5	Proper name	42	38	Medical care	3
6	Concept/content	38	39	Social control	3
7	Abstraction	32	40	Software	3
8	State/attribute	32	41	Substance	3
9	Event	30	42	Web	3
10	Group	25	43	Certainty	2
11	Place	25	44	Facility	2
12	Feeling	19	45	Force/strength	2
13	Entity	17	46	Game	2
14	Ability	12	47	Growth	2
15	Utensil	12	48	Clothing	1
16	Creation	11	49	Combustion	1
17	Body part	10	50	Commercial document	1
18	Commerce/exchange	9	51	Conformity	1
19	Part/portion	9	52	Cost	1
20	Explanation	7	53	Courage	1
21	Food/drink	7	54	Crystal	1
22	Measurement	7	55	Curiosity	1
23	Condition	6	56	Decision-making	1
24	Database	5	57	Duty	1
25	Difficulty	5	58	Epidemic disease	1
26	Equipment	5	59	Housing	1
27	Health issues	5	60	Imagination	1
28	Object	5	61	Influence	1
29	Possibility	5	62	Killing	1
30	Value	5	63	Metal	1
31	Category	4	64	Perception	1
32	Language unit	4	65	Title	1
33	Means of transport	4			
Total number of noun lexemes 664					

The creation of the Indonesian TreeTagger for use in LancsBox and CQPweb

Prihantoro
Universitas Diponegoro / Indonesia

Abstract – *TreeTagger* is a multilingual tagger capable of performing headword and POS tagging. However, before the completion of this project, Indonesian had not been supported. Thus, corpus query systems employing *TreeTagger* as a subsystem, such as *CQPweb* v.3.3.10 and *LancsBox* v.5, were incapable of annotating Indonesian texts. This context leads to the following research: 1) develop Indonesian language support for *TreeTagger*, 2) evaluate its performance, and 3) integrate the support into two popular corpus query systems, namely *CQPweb* and *LancsBox*, and demonstrate its functionalities. The research procedure can be concisely summarised as follows: training, annotation and evaluation, and incorporation. A pre-annotated corpus and lexicon were used in the training process. Headwords for the lexicon and corpus were semi-automatically added using *MorphInd*, augmented with expert revisions. The training produced an Indonesian *TreeTagger* parameter file, whose accuracy for POS and headword annotation was 96 per cent and 91 percent respectively. The parameter file has been incorporated into *LancsBox* v.6 and *CQPweb* 3.3.11, enabling support for the Indonesian language.

Keywords – *TreeTagger*; *CQPweb*; *LancsBox*; Indonesian; annotation

1. INTRODUCTION

Indonesian is the national and official language of Indonesia (Cohn and Ravindranath 2014: 134). It is a standardised Malay variety, spoken in Indonesia by more than 250 million speakers, as either a first or second language (Eberhard *et al.* 2022). This makes Indonesian the largest standardised Malay variety compared to other Southeast Asian Malay varieties. In light of this, the availability of corpus tools that support Indonesian is critical for the advancement of Indonesian corpus linguistics research. These tools should ideally be capable of performing at least three corpus-related tasks: 1) corpus data tokenisation, 2) annotation, and 3) analysis (Prihantoro 2022a), all of which are typically performed prior to interpreting corpus data findings.

Currently, a variety of computer programs are available to assist linguists with the three tasks mentioned above. For instance, a raw Indonesian corpus may be firstly tokenised using *Sastrawi* (Librian 2016) and annotated using *IPOSTagger* (Wicaksono



and Purwarianti 2010) or *MorphInd* (Larasati *et al.* 2011). The annotated corpus can then be imported into corpus query systems such as *AntConc* (Anthony 2024) or *WordSmith* (Scott 2024), among others, allowing users to query and implement basic to advanced corpus analysis techniques such as concordance, collocation, and keyword analyses.

Unfortunately, the preceding illustration demonstrates the potential difficulties faced by linguists who lack fundamental technical skills. First, multiple programs must be installed. Second, familiarity with the operation of all those programs and their transitions is required. In particular, *Sastrawi*, *IPOStagger*, and *MorphInd* are only accessible through the command line/terminal, whereas others, such as *AntConc* and *WordSmith*, are accessible through the Graphical User Interface (GUI).

CQPweb (Hardie 2012, 2023) and *LancsBox* (Brezina *et al.* 2018, 2020) are presented as improvements over the aforementioned programs because they allow all three of the tasks to be completed within a single user-friendly system. This enables linguists with limited technical proficiency to avoid the aforementioned complexity. While advanced corpus query functionalities are inherent to *CQPweb* and *LancsBox*, tokenisation and annotation tasks are accomplished by incorporating *TreeTagger* (henceforth TT; Schmid 1999, 2024), which is a tagger for many languages (English, French, German, and Chinese, among others) as a sub-system (i.e., third party software). Unfortunately, prior to completion of this project, an Indonesian TT had not been created, so neither *CQPweb* nor *LancsBox* could support annotation for Indonesian texts. This necessitated constructing a TT parameter file for Indonesian, which is referred to as the Indonesian TT. It can assist linguists who wish to annotate Indonesian texts with POS tags and headwords and implement corpus searches using these annotations in *CQPweb* and *LancsBox*.

In light of the previous discussion, the aims of the present study are:

- (1) to develop Indonesian language support for TT in the form of a TT parameter file;
- (2) to evaluate its performance;
- (3) and to integrate the parameter file into *CQPweb* and *LancsBox*, as well as demonstrating its functionality.

Once the performance-measured Indonesian TT has been developed and integrated into *CQPweb* and *LancsBox*, both systems will support Indonesian and may be used to

annotate a large corpus to their full capabilities. The present study therefore constitutes a practical contribution while, theoretically, it will allow researchers to carry out advanced quantitative and qualitative analysis by discovering new patterns, profiling texts, and conducting keyword analyses based on headwords or POS annotations, or a combination thereof.

2. LITERATURE REVIEW

2.1. *Why TreeTagger?*

In what follows, I review relevant literature pertaining to the first and second objectives in the study (see Section1). I argue that TT is a suitable system for the development of Indonesian language support, particularly headword and POS annotations. In the context of the Indonesian language, the term ‘headword’ refers to a monomorphemic word, not a lemma. For details, see Prihantoro (2021a, 2021b).

First, in terms of tokenisers, some tools have been developed: see, among others, *Sastrawi* (Librian 2016) or built-in tokenisers in NLP toolkits such, as the *Natural Language Toolkit* (NLTK; Bird *et al.* 2019) and *Spacy* (Vasiliev 2020). Nonetheless, some POS tagging systems such as *MorphInd* (Larasati *et al.* 2011) and TT (Schmid 1999) already include tokenisers. Here, I argue that the use of these systems is more efficient than the use of tokenisers in isolation.

As regards specific existing Indonesian POS taggers, there are several systems that can perform headword and POS tagging tasks on Indonesian texts, namely, the *Two-Level Morphological Analyser for Indonesian* (TLMA-Ind; Pisceldo *et al.* 2008), the *IPOSTagger* (Wicaksono and Purwarianti 2010), *MorphInd* (Larasati *et al.* 2011), and Indonesian POS taggers developed by (Rashel *et al.* 2014; Fu *et al.* 2018; and Maulana and Romadhony 2021).

TLMA-Ind is a morphological analyser and synthesiser. However, synthesis functionality is not required for this project. In terms of analysis, the system will output a headword and a POS tag, which may be followed by a morphological tag, when given an Indonesian word as input. Example (1) illustrates the output for the input word *memukul* ‘hit’ (Pisceldo *et al.* 2008: 149). The word is analysed by its headword *pukul* ‘hit’, shallow POS category (+Verb), and its morphological tag (+AV).

(1) *pukul* + Verb + AV

Nevertheless, the TLMA-Ind annotation scheme is extremely inadequate. It is limited to producing only four POS tags: noun, verb, adjective, and (literally) others. POS such as preposition, conjunction, etc. are included in the tag ‘others’. In addition, TLMA-Ind was evaluated against a list of words rather than a testbed corpus. This means that the words have been taken out of context.

In terms of its annotation scheme, the *IPOSTagger* is more advanced than TLMA-Ind. Unlike TLMA-Ind, which has only four POS tags, the *POSTagger* has 35 POS tags, including prepositions, interjections, and numerals, among others. These tags are provided alongside word tokens delimited by forward slashes. In contrast to TLMA-Ind, the output is unambiguous. A word token can only have a single analysis. In terms of performance, the *IPOSTagger* achieved 99 per cent (the best) accuracy when evaluated against a testbed corpus. However, it only outputs POS annotation, which may be considered a drawback (headword annotation is not included). See, for instance, the sample output in (2), obtained from the following input sentence *saya berdiri di jalan* ‘I stood on the road’.

(2) *saya/PRP berdiri/VBI di/IN jalan/NN*
I/PRONOUN stood/INTRANSITIVE_VERB on/PREPOSITION road/NOUN
 ‘I stood on the road.’

Headword annotation appears to be understudied. The documentation for the *IPOSTagger*, *MorphInd*, and several other Indonesian POS taggers mentioned above does not report headword annotation evaluation. I would like to bridge this gap by reporting both headword and POS tagging accuracy.

MorphInd is claimed to be a morphological analyser, but it actually includes headword and POS annotation functions. For instance, it analyses *mengirimkan* ‘send something’ as a word composed of three morphemes (presented as underlying instead of surface forms), *meN+*, *irim<v>*, and *+kan*. The headword is indicated by a headword POS (here, *<v>*). The word POS is presented at the end, here *_VSA* (serb singular active), whose full word analysis is shown (3) and Table 1, below. Although decomposing words into morphemes is one of *MorphInd*’s strengths, identifying morphemes other than the headword is not the focus of this project.

(3) *meN+irim<v>+kan_VSA*

Output	Description
meN	Underlying form of the active voice prefix
+ kirim	Headword of <i>mengirimkan</i>
<v>	Headword POS
+kan	Underlying form of the suffix*
_VSA	Word tag

Table 1: Morphological annotation description for example (3)¹

Unlike other aforementioned systems, *MorphInd* has been widely used by scholars (Chung and Shih 2019; Denistia 2023). In its official report (Larasati *et al.* 2011: 119), the performance of *MorphInd* was measured in terms of its overall token coverage of 84 per cent. However, scholars report different evaluative measures. Instead of coverage, Denistia (2023: 15) and Prihantoro (2021a: 175) reported that *MorphInd*'s accuracy was measured at 84 per cent and 89 per cent, respectively, which appears to be the result of the different testbed corpora employed. While Denistia (2023) reported 84 per cent accuracy, her evaluation metric is not coverage and therefore differs from the evaluation metric used in *MorphInd*.² Regardless, the output of *MorphInd* is quite complex. As Larasati *et al.* (2011: 122) mention, the *MorphInd* tagset is influenced by the *Penn Treebank* tagset.³ Thus, it does not seem to fully reflect Indonesian morphosyntax.⁴

Unlike the foregoing systems, taggers developed by Fu *et al.* (2018) and Maulana and Romadhony (2021) are not publicly available. Rashel *et al.*'s (2014) demo tagger is claimed to be available but is in fact inaccessible. Thus, the present review of literature depends solely on their papers, whose linguistic content is very limited. The accuracy of these three taggers is reported at 96 per cent, 79 per cent, and 92 per cent respectively. As for the tagset, the number of tags in Rashel *et al.*'s (2014) and Fu *et al.*'s (2018) tagsets are 23 and 29 tags, respectively, but in Maulana and Romadhony's (2021) tagger, the number of tags is not mentioned. Before the project was completed, TT did not support Indonesian at all. As the tagset and its performance evaluation were still unknown at the time of review, the status of tagset and evaluation is yet unknown before the conclusion of this project. This is shown in Table 2.

¹ Note both surface and underlying forms are identical for *-kan*.

² For a qualitative evaluation of *MorphInd*, see Chung and Shih (2019) and Prihantoro (2021b).

³ <https://www.sketchengine.eu/penn-treebank-tagset/>

⁴ For additional criticism of the *MorphInd* tagset, see Prihantoro (2021a).

Tagger	Tagset	Evaluation	Access	Operation	Design
<i>TLMA-Ind</i>	4	Precision (89%)	Accessible	Terminal	RB
<i>IPOSTagger</i>	25	Accuracy (99%)	Accessible	Terminal	DD
<i>MorphInd</i>	39	Coverage (84%)	Accessible	Terminal	RB+DD
Rashel <i>et al.</i> (2014)	23	Accuracy (79%)	Accessible	Terminal	RB
Fu <i>et al.</i> (2018)	29	Accuracy (96%)	NM	Terminal	DD
Maulana and Romadhony (2021)	NM	Accuracy (92%)	NM	Terminal	DD
<i>TreeTagger</i>	UNK	UNK	Accessible	Terminal	DD

Table 2. Summary of taggers reviewed in this project (NM=Not mentioned, RB=Rule-Based, DD=Data-Driven, UNK=unknown yet as this project was not concluded)

As for the design of a tagger, Table 2 suggests that there are at least two designs: 1) rule-based (or linguistic) and 2) data-driven (or statistical) approaches (a third design is a combination of the two). The differences between these two approaches are concisely described in (Voutilainen 1999: 9–10). Regardless of the fundamental methodological differences, (Prihantoro 2021a: 300) states that, depending on the quality of the resources, the two methods may generate taggers with relatively similar performance.

In terms of evaluation, Table 2 suggests that there are at least three evaluative measures: coverage, precision, and accuracy. I exclude ‘coverage’ here because it only measures the proportion of tokens a tagger recognises correctly, not tokens corresponding to headword or POS tags. Prihantoro (2021a: 242–243) argues that precision suits taggers whose output may be ambiguous, while accuracy suits taggers whose output is unambiguous. Thus, the preference would depend on the characteristics of the output.

Note that all the aforementioned systems can potentially be challenging when users are not skilled in programming, at least at the basic level, both for program installation as well as their operation. *MorphInd*, for instance, can only be installed on Mac or Unix-like distributions, such as Ubuntu. The Linux subsystem must be installed if it is to be utilised on a Windows machine. Then, subsequent applications must be pre-installed: *SVN*, *Perl*, *Foma*, and *Subversion*. Operation is performed through a command line or terminal application. TT is no exception. Example (4) illustrates a TT command line to annotate the source.txt file and place the output in the result.txt file.

(4) tree-tagger.exe english.par source.txt result.txt –token –lemma

Note that source.txt requires a format of one token per line, which can be implemented using an existing *Perl* tokeniser script in the TT system. This regrettably adds another layer of complexity, as the source file must be tokenised before applying the tagging

command line. Table 3 shows the difference between input and output file formats. In the latter, headword and POS annotation are added.

source.txt		result.txt	
Tagger	Token	HW	Tag
There	There	there	EX
are	are	be	VBR
some	some	some	DD
apples	apples	apple	NN2

Table 3: Input and output format (CLAWS-7 tagset, HW=headword)

A further complication is that the output file must be reformatted based on the format accepted by the preferred corpus query system. Example (5) shows a format accepted by *AntConc*, which does not include headword annotation. *AntConc* mandates that the list of lemmas must be maintained in a separate file. Consequently, tokens and lemmas must be extracted and stored as a lemma list.

(5) there/EX are/VBR some/DD apples/NN2
 there/EXISTENTIAL_THERE are/ARE some/DETERMINER
 apples/PLURAL COMMON NOUN

To summarise, I argue that it is more effective to develop the Indonesian TT from the ground up, despite the complexities that it presents. This is due to the fact that the TT has been integrated as a subsystem of several corpus query systems, including *CQPweb* and *LancsBox*. Once the Indonesian TT is incorporated into *LancsBox* and *CQPweb*, end users will be able to bypass all of these complex processes by simply clicking a button to perform POS and headword annotation, as well as Indonesian corpora searches and other advanced functionalities (collocation, keyword, etc.) based on this annotation. While the TT's accuracy was routinely measured at between 94 per cent and 96 per cent (Schmid 1999), it should be noted that at the time the project was carried out, it was unknown to what extent the accuracy was consistent for Indonesian. Evaluation methods will be explained in more detail in Section 3, and their implementation will be described in Section 4. As with other languages, TT can perform headword as well POS annotation and can be used to analyse more than 20 languages, excluding Indonesian, prior to the completion of this project.

2.1. Why CQPweb and LancsBox?

In relation to the third objective of this project, I argue that *CQPweb* and *LancsBox* are two state-of-the-art sophisticated corpus query systems that are suitable for this project. Note that other query systems are also in use, including *Sketch Engine* (Kilgarriff *et al.* 2014) and *English-Corpora* (Davies 2024). Other popular corpus tools are *AntConc* and *WordSmith*. These are the main corpus linguistic tools surveyed by Gomide (2020).

These systems can be divided into two major categories: web-based and desktop applications (or third- and fourth-generation corpus linguistic tools as shown in Gomide (2020: 24–26). *CQPweb*, *Sketch Engine*, and *English-Corpora* applications are web-based. Users are not required to install anything on their computers. These applications are accessible via a variety of internet browsers, including *Google Chrome*, *Safari*, or *Firefox*, among others, regardless of the operating system (*Windows*, *Mac*, or *UNIX* distribution). Conversely, *LancsBox*, *WordSmith*, and *AntConc* are computer applications which users must install. Every system has its own specifications. For instance, *WordSmith* cannot be installed on *Mac* or *UNIX*-like operating systems. Users are therefore required to install additional software, such as *Wine* or *Parallel*.

All of the aforementioned systems are user-friendly in that they do not require users to be able to code or program using languages such as *Bash*, *Python*, *PHP*, *Perl*, and the *Windows* command line, among others. Only three of the aforementioned systems are capable of performing POS and headword annotation: *Sketch Engine*, *CQPweb*, and *LancsBox*. These are the candidates for the incorporation of the Indonesian TT. The three systems have integrated TT into their operations. *Sketch Engine*, however, is excluded because it requires backend access to incorporate the Indonesian TT, which I do not have access to. Moreover, *Sketch Engine* is a proprietary system, while *CQPweb* and *LancsBox* are open-source systems.

Regarding *LancsBox*, there are two ways to implement the Indonesian TT. The first is from the server, which I do not have access to, and the second is from the configuration of the resource file, which is freely accessible to users once downloaded to a computer. Therefore, *LancsBox* is preferred. *CQPweb* is similar to *Sketch Engine* in that developer access is required, but the author has access to *CQPweb* admin control. Thus, the primary reasons for selecting *CQPweb* and *LancsBox* are that they are non-commercial and open-access systems. In addition, *CQPweb* and *LancsBox* represent two distinct types of advanced corpus query systems: web-based and desktop applications. Note that *LancsBox*

here is different from *LancsBox X* (Brezina and Platt 2024), another variant of *LancsBox* which uses *Spacy* instead of TT for POS tagging purposes.

3. METHODOLOGY

The research procedure in this project is presented in chronological order, with emphasis on sections pertinent to the research objectives. The procedure can be divided into four phases: 1) training, 2) annotation and 3) evaluation (whose process is shown in Figure 1), and, finally, 4) incorporation.

In accordance with the TT framework, I created the Indonesian TT (the TT parameter file used for tagging Indonesian texts) using a data-driven approach (see Section 2.1), which requires a training process. First, I prepared a number of resources to train the TT and output a parameter file (language support for the TT), as prescribed by Schmid (1999, 2024). The resources are a training corpus, a lexicon, a file containing tags for guessing unknown words (referred to as open-class file), and a TT training application.

I chose a 250,000-word pre-annotated Indonesian corpus ([1] in Figure 1) made available by Dinakaramani *et al.* (2014) to ensure the quality of the training corpus. The corpus was manually annotated and checked using its own tagset (which I will refer to here as the UI tagset), whose creation is discussed in Dinakaramani *et al.* (2014).

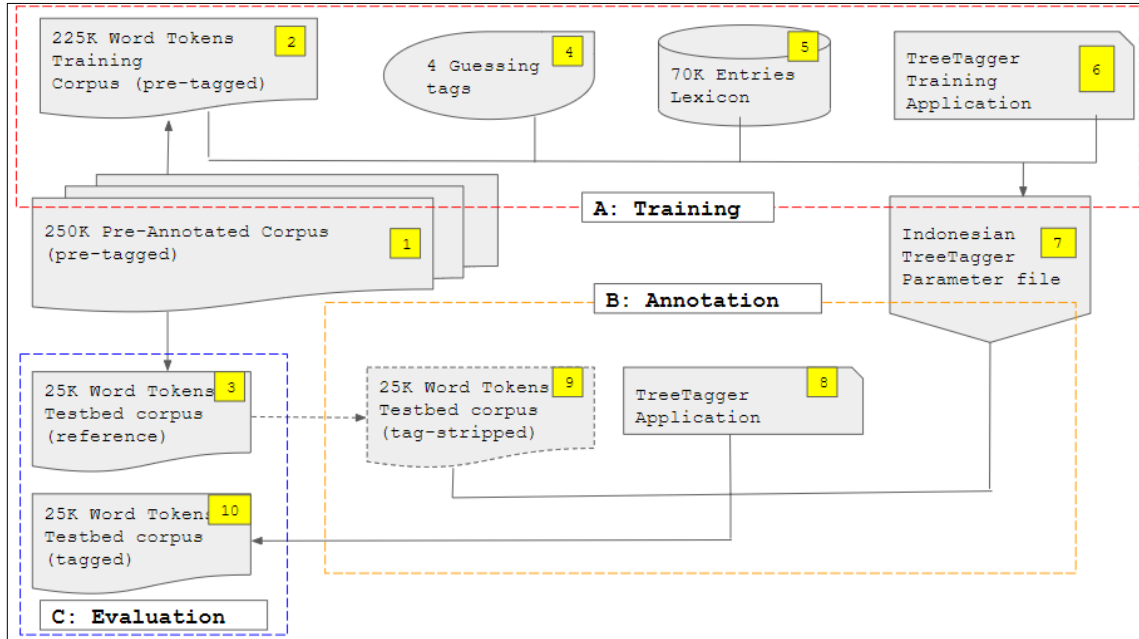


Figure 1: Training, annotation, and evaluation phases

The corpus was then divided in two parts. The first 25,000 words ([3] in Figure 1) were extracted as a reference testbed corpus, which was used in the final stage to evaluate the application's performance in the evaluation phase.

The remaining words, 225,000k ([2] in Figure 1), served as the training corpus, a resource which was processed by the TT training application ([6] in Figure 1) to create the Indonesian TT ([7] in Figure 1) in the training phase. In addition to the training corpus, another resource required to create the Indonesian TT is the lexicon, a file containing a list of words with the corresponding tags, and if possible, headwords ([5] in Figure 1), and open ([4] in Figure 1) class file. The lexicon was compiled by adapting entries from *The Great Indonesian Dictionary* (Bahasa 2005), or abbreviated as KBBI III, to TT format. In terms of the number of words, the lexicon (around 70,000 words) is just around 30 per cent of the training corpus. However, it is larger in terms of coverage because all words are unique (different from the training corpus in which the same words may be repeated). The lexicon is expected to improve the coverage of the Indonesian TT.

The open class file consists of possible tags for unknown words to anticipate paucity in the training corpus and lexicon. If a word tag (and headword) is still unknown by the Indonesian TT, even after consulting the corpus and lexicon, the program will use orthographic cues to supply a most likely tag. For instance, when an unknown Indonesian word begins with *pe-* and ends with *-an*, the most likely tag is NN because, in the training corpus and lexicon, the majority of words with the same orthographic context are tagged NN. To prevent the program from tagging the unknown word using closed class tags, the list of the content word tags was made explicit in the open class file.

I inspected the words in the testbed corpus against the words in the corpus and lexicon. The output was a list of words in the testbed corpus that were not present in the training and the lexicon. There were no function words (such as prepositions or conjunctions), but all the words were content words (nouns, verbs, adjectives, adverbs). Therefore, five content word tags were selected for guessing, namely, NN (noun), VB (verb), JJ (adjective), RB (adverb) and NNP (proper noun), ruling out the probability of these unknown words to be categorised into function words.

Once all the components were complete, I began the training phase, which aimed to create the Indonesian TT, the parameter file, and a language model file for tagging Indonesian texts. The training ([A:Training] in Figure 1) was administered within a TT environment using the abovementioned resources (training corpus, lexicon, guessing tags,

and the TT training application: an application used to create a parameter file for tagging from the aforementioned resources). The training for the POS tags and headwords is an integrated process, not a separate one. When the TT training application was applied to the training corpus and lexicon, it read all information (token, POS tag, headword) to build a language model. The result was an Indonesian TT parameter file ([7] in Figure 1). This concluded the training phase.

The performance of the Indonesian TT parameter file was then evaluated in the two next phases: annotation and evaluation. In the annotation phase, the parameter file was first applied to the testbed that had its tags and headwords stripped ([9] in Figure 1) using TT application ([8] in Figure 1). This concluded the annotation phase,

In the evaluation phase, the output ([10] in Figure 1) was compared to the headword and POS tags of the reference testbed ([3] in Figure 1). Performance evaluation is expressed in (%) accuracy: the percentage of tags and headwords correctly tagged. This is because, by default, the TT does not produce ambiguous output.

In addition, as demonstrated by the evaluation of taggers in different languages such as Tamil (Thavareesan and Mahesan 2020), Korean (Park and Seo 2015), and Turkish (Can *et al.* 2017), among many others, accuracy is a common metric for measuring POS tagger performance. Evaluation is relevant to the second objective, which is to assess the performance of Indonesian TT. This concludes the evaluation phase.

The last phase was the incorporation of the Indonesian TT into *LancsBox* and *CQPweb*. I will demonstrate how to include the parameter file in the configuration of the *LancsBox* resource file so that it can be used to tag Indonesian texts. As for *CQPweb*, I will demonstrate how users can utilise the ‘install your own corpus’ functionality, which allows them to select the Indonesian TT as the language support. This is relevant to the third objective of this project.

4. RESULTS AND DISCUSSION

4.1. The Indonesian TT

The construction of the Indonesian TT is described in this section. The creation of TT language support requires four resources: 1) a training corpus, 2) a lexicon, 3) a guesser list, and 4) a tree-tagger training application (already provided in the TT environment).

As for the creation of the training corpus, the pre-annotated corpus (Dinakaramani *et al.* 2014) was downloaded from Fam Rashel’s Github repository (Rashel 2016). The file was saved as training.txt. As mentioned in Section 3, the first 25,000 words were extracted and used to create a testbed corpus. This extraction was extended to line 25020 (see Table 4) to guarantee that the final sentence was completely extracted. The resulting reference testbed was named ‘testbed-reference.txt’.

Line number	Token	Gloss
25001	<i>Pemerintah</i>	‘Government’
25002	<i>Juga</i>	‘Also’
25003	<i>Melakukan</i>	‘Do’
...
...
25017	<i>Untuk</i>	‘To’
25018	<i>Meningkatan</i>	‘Improve’
25019	<i>Kesiagaan</i>	‘Readiness’
25020	.	.

Table 4: Extension to 25,020 tokens for the testbed corpus

As for the lexicon, entries were obtained from KBBI III, downloaded from the *Kateglo* repository (Lanin *et al.* 2019). These are all full-form entries that are linguistically more comprehensive than root-form lexicons (Prihantoro 2022b) and vocabulary indexes (such as Lun *et al.* (2023)). Several modifications to the lexicon format and content were implemented. First, information other than entries and their respective POS tags was removed. The tagset used in the lexicon had not been adapted to match the tagset used in the training corpus.

I chose to adapt the KBBI tagset (the tagset used in the lexicon) to form the UI tagset (Dinakaramani *et al.* 2014), because the latter is more fine-grained than the KBBI tagset. The KBBI tagset only consists of noun, verb, adjective, adverb, numeral, and *kata tugas*, which is a bin tag for other tags not yet covered (prepositions and conjunctions, among others), thus can simply translate to the tag ‘others’. Ordinal and cardinal numerals, for instance, are subsumed under numerals in KBBI using the tag *num*. However, if a comparison is made Larasati *et al.*’s (2011) tagset is more fine-grained than the UI tagset. Prihantoro (2021b) argues that some of Larasati’s tags are incompatible with Indonesian morphosyntax. For example, VSA stands for verb singular active. The incorporation of ‘singular’ appears to have been influenced by English grammar, in which verbs are marked according to their subject (singular or plural). The tagset developed by

Wicaksono and Purwarianti (2010) is slightly more refined than the UI tagset. For instance, it distinguishes between transitive and intransitive verbs. Nonetheless, transitivity must be determined at the level of syntax, which is beyond the scope of this study.

After selecting the tagset, I moved on to tagset adaptation. When UI tags were more specific, I manually revised the POS tags. For instance, the tag *num* for *pertama* ‘first’ was converted into OD (OrDinal number), while *satu* ‘one’ was converted into CD (CarDinal number). In some instances, a simple find-replace operation could be used to effectively adapt the lexicon. For instance, the verb tag *v* in KBBI could simply be replaced with the VB (VerB) tag.

The absence of headwords in the lexicon was an issue because TT requires the token-tag-headword sequence to be present in each entry line (each one delineated by a tab), as shown in Table 5. Ambiguous entries (e.g., *bisa* as a modal verb ‘able to’ and also as a noun ‘venom’) must be on the same line. Following the tag-headword pair for the first meaning, another tag-headword pair for the second meaning must be included in the same line. This differs from the *SANTI-morf* lexicon (Prihantoro 2022b) in which ambiguous entries must be placed on separate lines.

Type	Entry	Tag1	HW1	Gloss1	Tag2	HW2	Gloss2
Unambiguous	<i>Mengambil</i>	VB	<i>Ambil</i>				
Ambiguous	<i>Bisa</i>	MD	<i>Bisa</i>	‘Able to’	NN	<i>Bisa</i>	‘Venom’

Table 5: Unambiguous and ambiguous entry format (HW=headword)

To ensure that all lexical items in the lexicon were associated with their corresponding headwords, I applied *MorphInd*. Note that *MorphInd*’s output (as shown in Section 2.1) includes affixes, headword POS and full word tags. These elements were omitted to ensure the lexicon was formatted correctly.

To ensure that every lexical item in the training corpus is present in the lexicon, the content of the training corpus training.txt was incorporated into the lexicon. Similar to the initial version of the lexicon discussed earlier, the corpus lacked a headword; only lexical items and their corresponding POS tags were included. I added a headword to each lexical item using *MorphInd* to make the content compliant with the lexicon format. *MorphInd* managed to supply 98 per cent of the lexical items’ corresponding headwords. Out of these, 92 per cent of the headwords were supplied correctly. I then manually revised the

analyses until they were 100 per cent accurate and added the missing headwords. Table 6 illustrates some of the lexical items whose headwords were incorrectly supplied.

Lexical item	Gloss	Incorrect HW	HW Corrected
<i>Menurut</i>	‘According to’	<i>Menurut</i>	<i>Turut</i>
<i>Menjelang</i>	‘Near’	<i>Menjelang</i>	<i>Jelang</i>
<i>Menarik</i>	‘Pull’	<i>Menarik</i>	<i>Tarik</i>
<i>Pesaing</i>	‘Competitor’	<i>Pesaing</i>	<i>Saing</i>
<i>Mendatang</i>	‘Next’	<i>Mendatang</i>	<i>Datang</i>

Table 6. Sample of incorrect headwords (HW=headword)

The next step was to remove duplicates because some of the original lexical items in the lexicon were identical to lexical items obtained from the corpus. The final lexicon output was saved as *lexicon.txt*.

As for the third element, guessing tags, only four tags were used and all were content word tags, namely, VB, NN, JJ, NNP (verbs, nouns, adjectives, and proper nouns). These tags were stored in *guess.txt*.

All the resource files (*lexicon=lexicon.txt*, list of tags for guessing=*guess.txt*, and training corpus=*train.txt*) required to build an Indonesian parameter file for TT have been completed. Training was conducted in a TT environment using a *training-tree-tagger.exe* file. The command line below executed the training process, resulting in a parameter file called ‘*indonesian23.par*’, as shown in (6).

(6) *train-tree-tagger.exe lexicon.txt guess.txt train.txt indonesian23.par*

The creation of Indonesian TT (Figure 2) finalised the training process, ensuring that the first objective of this paper was accomplished. The next step will be discussed in what follows and deals with annotation and evaluation.

```
C:\Users\priha\TreeTagger>train-tree-tagger.exe lexicon.txt
guess.txt training.txt indonesian23.par

train-tree-tagger -cl 2 -dtg 0.50 -sw 1.00 -ecw 0.15 -atg 1
.20 lexicon.txt guess.txt training.txt indonesian23.par

    reading the lexicon ...
        reading the tagset ...
        reading the lemmas ...
        reading the entries ...
        sorting the lexicon ...
        reading the open class tags ...
    calculating tag frequencies ...
finished.
    making decision tree ...
164    saving parameters ...

Number of nodes: 165
Max. path length: 29
```

Figure 2: An Indonesian TT parameter file created

4.2. Annotation and Evaluation

In this section, I will discuss the second objective of this project: namely, evaluating the performance of the Indonesian TT. First, a copy of the testbed-reference.txt was created and renamed as ‘testbed-stripped.txt’. The latter is the actual testbed. Next, all tags in the actual testbed-stripped.txt were removed. Subsequently, both the TT application and the Indonesian parameter file were used to tag the actual testbed. The command line that executed the annotation can be seen in (7).

(7) tree-tagger.exe indonesian23.par testbed-stripped.txt testbed-tagged.txt -token -lemma

The evaluation was conducted by comparing the reference testbed’s annotations, which were regarded as the gold standard, with the annotations obtained from the Indonesian TT (the resulting file was testbed-stripped.txt) whose accuracy measure is defined as follows: (correctly annotated tokens/all tokens) * 100 = accuracy (%). Tokens are considered to be correctly annotated only if they contain matching annotations. Table 7 illustrates 100 per cent accuracy for both POS and headwords for five tokens, because annotations from the actual testbed match all annotations from the reference testbed.

Gloss	Token	RT		AT		TM	HWM
		Token	Tag	Token	Tag		
‘Rise’	<i>Naik</i>	<i>Naik</i>	VB	<i>Naik</i>	VB	Yes	Yes
‘From’	<i>Dari</i>	<i>Dari</i>	IN	<i>Dari</i>	IN	Yes	Yes
‘IDR’	<i>Rp</i>	<i>Rp</i>	SYM	<i>Rp</i>	SYM	Yes	Yes
‘50’	<i>50</i>	<i>50</i>	CD	<i>50</i>	CD	Yes	Yes
‘Become’	<i>Menjadi</i>	<i>Jadi</i>	VB	<i>Jadi</i>	VB	Yes	Yes

Table 7: Matching segments (RT=Reference testbed, AT=Actual testbed, TM=Tag match HWM=headword match)

When compared to previous POS tagger studies, this is the first study that measures the performance of headword annotation. In studies on the Indonesian POS tagger reported in Section 2 (Wicaksono and Purwarianti 2010; Larasati *et al.* 2011; Rashel *et al.* 2014; Fu *et al.* 2018; Maulana and Romadhony 2021) headword annotation is not measured. The fact that the accuracy of headword annotation is left unmeasured, however, seems to be a common phenomenon in studies reporting the creation of taggers for other languages such as Tamil (Thavareesan and Mahesan 2020), Korean (Park and Seo 2015), and Turkish (Can *et al.* 2017). Prihantoro (2021a) evaluated the accuracy of *MorphInd*’s

headword annotation (89%). Nevertheless, that study is about a morphological annotation system, not about a POS tagger.

One of the possibilities why the accuracy of headword annotation does not exceed 94 per cent is because TT does not know how to lemmatise proper nouns that it has not seen in training. Except for proper nouns in the training corpus that have been lemmatised, all proper nouns are marked as ‘unknown’. For instance, there are six *Mayapada* ‘the name of a private bank’ in the testbed corpus, but none in the training corpus. As a result, all of the headword annotations for each token are ‘unknown’, and thus do not match. Instead of giving ‘unknown’, it might be best to copy the word as a whole into the headword slot. Note that in a corpus query, it is less likely for a user to search for a proper noun by looking at the headword. The rate of incorrect headword-annotated by POS categories can be observed in Table 8.

Word class	Percentage	Example
Proper noun	33	<i>Mayapada, Adaro, Xenia</i>
Common noun	25	<i>Perkemahan</i> ‘camping ground’, <i>kekhalifahan</i> ‘caliphate’, <i>peremasan</i> ‘squeezing process’
Verb	27	<i>Menjelang</i> ‘slightly before’, <i>menorehkan</i> ‘inscribe’, <i>melempari</i> ‘throw something repeatedly’
Adverb	8	<i>Secepatnya</i> ‘as soon as possible’, <i>sejujurnya</i> ‘honestly’, <i>semestinya</i> ‘as it must have’
Adjective	7	<i>Terendah</i> ‘lowest’, <i>seimbang</i> ‘ugliest’, <i>terjelek</i> ‘worst’

Table 8: Rate of incorrect headword annotation by POS categories

Another cause of errors is the paucity in the training corpus or lexicon (or both), whether missing entries or inaccurate lemmatisation. For instance, the headword of *berkeliaran* ‘wander’ is *keliar*. However, in the lexicon and the training corpus, the headword of *berkeliaran* is *berkeliaran*. This was overlooked during my earlier inspection, and consequently, TT made this error. In the future, a more thorough examination will be conducted.

In terms of POS tagging accuracy, the Indonesian TT was only three per cent less than Wicaksono and Purwarianti’s (2010) tagger, but four per cent better than the most recent Indonesian tagger developed by Maulana and Romadhony (2021). Nonetheless, I argue that a better comparison should be made using Rashel *et al.*’s (2014) study, as we employ the same tagset and training corpus. Here, the Indonesian TT performed 17 per cent better than Rashel *et al.*’s (2014) tagging experiment. This is summarised in Table

9. Note that both TT and Rashel *et al.*'s (2014) taggers differ slightly in terms of tagger design; the former adopts a data-driven approach, while the latter is rule-based. I suspect that the linguistic rules embedded in Rashel *et al.*'s (2014) resources may not be detailed enough, thus resulting in lower performance.

Tagger	POS Tagging	Headword
Rashel <i>et al</i> (2014)	79	NA
Wicaksono and Purwarianti (2010)	99	NA
Maulana and Romadhony (2020)	92	NA
<i>MorphInd</i> (2011)	89	89
Fu <i>et al.</i> (2018)	95	NA
The Indonesian <i>TreeTagger</i>	96	91

Table 9: Headword and POS tagging accuracy percentages

I then conducted further tests by compiling data for five alternative testbed corpora from texts not included in the corpus that derives the training corpus. These texts were created by extracting more than 25,000-word token texts from the Indonesian corpus data in the *Leipzig Corpora Collection* (Quasthoff *et al.* 2014) from 2016 to 2020.

As shown in Table 10, it turns out that POS tagging performance varies between 92 per cent and 96 per cent, whereas headword annotation performance varies between 87 per cent and 94 per cent. The mean values for POS and headword annotation are 94 per cent and 91 per cent respectively. The precision of POS tagging remains above the threshold established by Schmid (1999), whereas the precision of headword annotation is below it. That the accuracy of headword annotation is always lower than POS tagging requires further examination. One possibility for this to happen is the nasalisation rule in Indonesian when affixation takes place. The morphophonological alternation invoked by this nasalisation rule translates to orthography (Prihantoro 2021a), causing some forms to be failed to be reduced into the correct headword forms to consonant alternation or deletion (e.g., *men* 'ACV' + *tanam* 'plant' > *men* +[t]an~~a~~m = *menanam* 'plant'). In testbed version 4, their accuracy is identical. This is an outlier as compared to other versions, but similar to the findings in Prihantoro's (2021a). Nonetheless, the performance evaluation of the Indonesian TT has been completed.

Alternative Testbed	Pos Tagging Accuracy	Headword Tagging Accuracy
Version 1	96	94
Version 2	92	87
Version 3	92	92
Version 4	95	91
Version 5	94	91
Mean	94	91

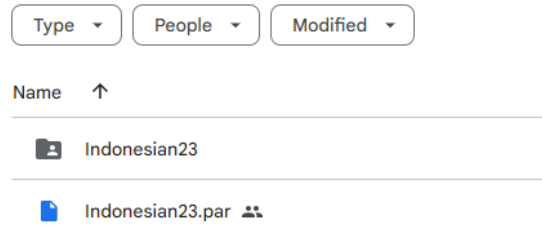
Table 10: Headword and POS tagging accuracy percentages

4.3. *LancsBox v.6.0 and CQPweb v.3.3.11*

LancsBox v.4 (Brezina and Platt 2024) and v.5 (Brezina *et al.* 2020) lacked Indonesian language support, as did *CQPweb* 3.3.10 and earlier versions. Regarding the third aim in the study, in what follows, I demonstrate that support for the Indonesian language has been added to *LancsBox* version 3.6 and *CQPweb* version 3.3.11.

4.3.1. *LancsBox v.6.0*

First and foremost, ensure that all required resource files have been downloaded. As pointed out in Section 4.1, the first file required is the Indonesian TT parameter file (indonesian23.par). Other *LancsBox* specific resource files, such as a list of acronyms, are stored in a folder named ‘Indonesian23’. These files are accessible from this repository (<https://tinyurl.com/indonesian23>), as shown in Figure 3.

Figure 3: The Indonesian language support resource files for *LancsBox*

Second, ensure that *LancsBox* v.6 is the version of *LancsBox* installed on the computer. Next, locate the *#LancsBox* folder. In *Windows*, it is accessible via the private user folder on the C disc (C: Users). If *LancsBox* is properly installed, the *#LancsBox* folder should be at the top (Figure 4).

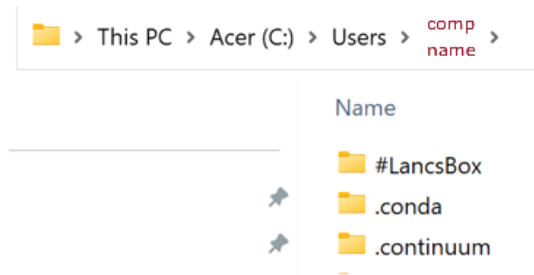


Figure 4: Location of *LancsBox* folder in *Windows*

From the *#LancsBox* folder, go to *Resources > tagger > models* (*C:\Users\comp_name\#LancsBox\resources\tagger\models*). The *models* folder contains the language parameter files supported by *LancsBox*. By default, the English TT parameter file (*english-utf8.par*) will be displayed. If *LancsBox* has been used to analyse other languages, such as Chinese and French, the parameter files for those languages will also be visible (*chinese.par* and *french.par*). Then, place the Indonesian parameter (*indonesian23.par*) file alongside other parameter files (Figure 5).

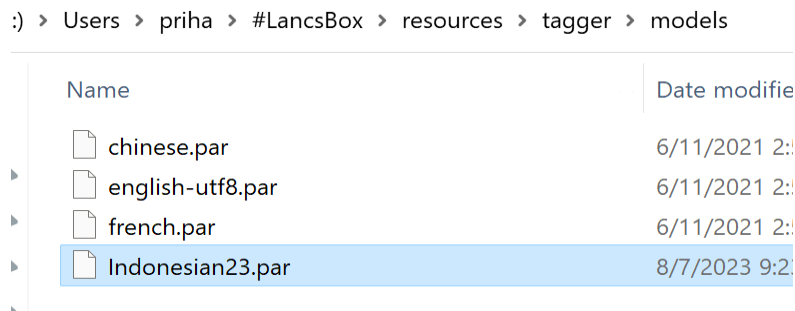


Figure 5: Parameter files folder in *LancsBox*

Returning to the *Resources* folder, navigate to the *Languages* folder (*Resources > Languages*). Folders containing languages for which support has been enabled will be visible. For example, the default language will be English. However, if you have activated support for additional languages, the folders for those languages will also be displayed. Insert the *Indonesian23* folder here (Figure 6).

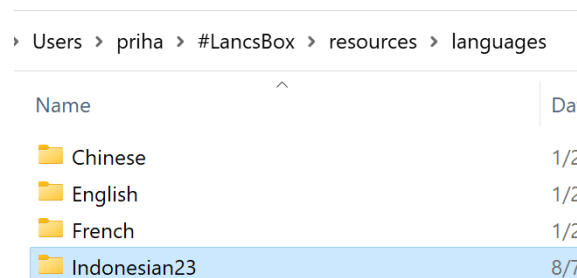


Figure 6: Language resource folder in *LancsBox*

Finally, start *LancsBox*. If you have started already, close *LancsBox* and start again. Once you have restarted, go to Language. Click on the triangle button next to English and select Indonesian23 to enable support for the Indonesian language (Figure 7). From this point onwards, use *LancsBox* normally.

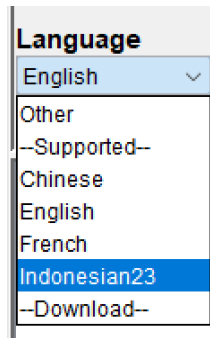


Figure 7: Selecting the Indonesian language support

As headword and POS tags have been incorporated, it is now possible to search the annotated Indonesian corpora using lemma and POS tags. In KWIC (Figure 8), for instance, users may enter POS tags or keywords in the query box. The list of tags is accessible via the Tags button within Import Options (Figure 9).



Figure 8: KWIC query box in *LancsBox*

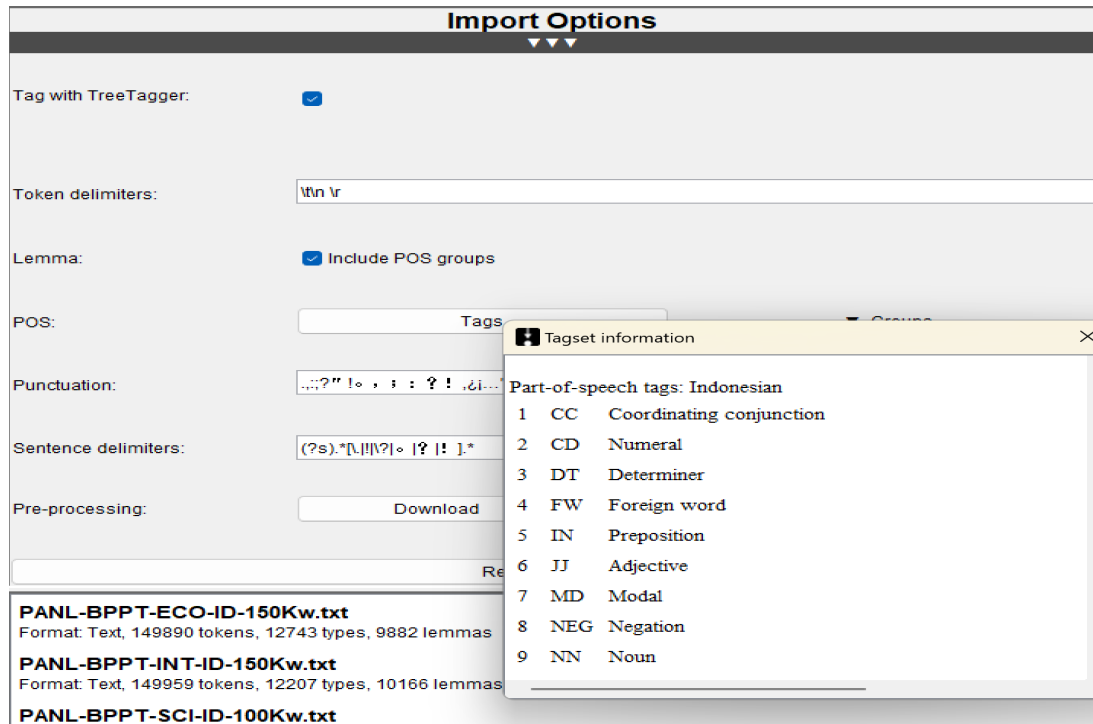


Figure 9: Indonesian tagset information in *LancsBox*

In addition to KWIC, POS and headword filters can also be used in other *LancsBox* tools. For example, in *GraphColl*, we can filter a collocation analysis result by POS and/or lemma. Figure 10 shows adjective collocates of *gol* ‘goal’ in the *BPPT-PAN* corpus indexed in *LancsBox*, namely, *imbang* ‘draw’, *tunggal* ‘sole’, *banyak* ‘many’, and *unggul* ‘excellent’.

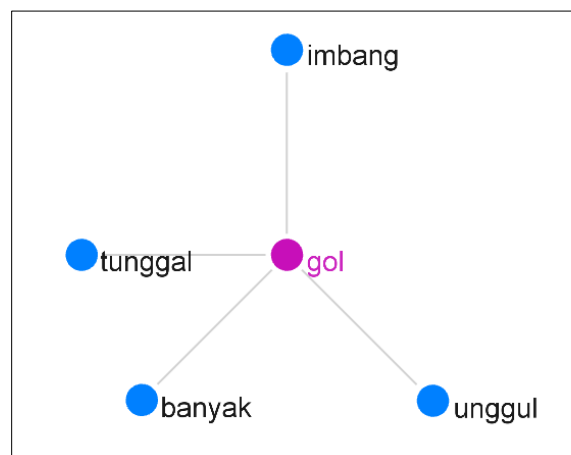


Figure 10: Sample of adjective collocates of *gol* in the *BPPT-PAN* corpus in *LancsBox* (LR=5, Stat=Dice, Threshold=Default, View=Lemma, Filter=adjective)

Figure 11 is a *GraphColl* showing collocates from three words of the same headwords: *mencetak* ‘print/to score’, *pencetak* ‘printer/scorer’, and *dicetak* ‘be scored/printed’. We

can observe that *gol* ‘goal’ is a shared collocate of *pencetak* ‘scorer/printer’ and *mencetak* ‘score/print’, but not of *dicetak* ‘be printed/scored’. The relativiser *yang* ‘who/that’ is a shared collocate of *mencetak* ‘score/print’ and *dicetak* ‘be scored/printed’ (active and passive forms). The remaining collocates are exclusive (not shared). For instance, *terbanyak* ‘highest quantity’ is an exclusive collocate for *pencetak* ‘scorer/printer’. Likewise, *penalti* ‘penalty’, and other 43 types (densely populated) are collocates for *mencetak* ‘score/print’. The overpopulated collocates from *mencetak* ‘print/score’ were preserved, because if the collocation parameters are changed, collocates from other nodes might be hidden.

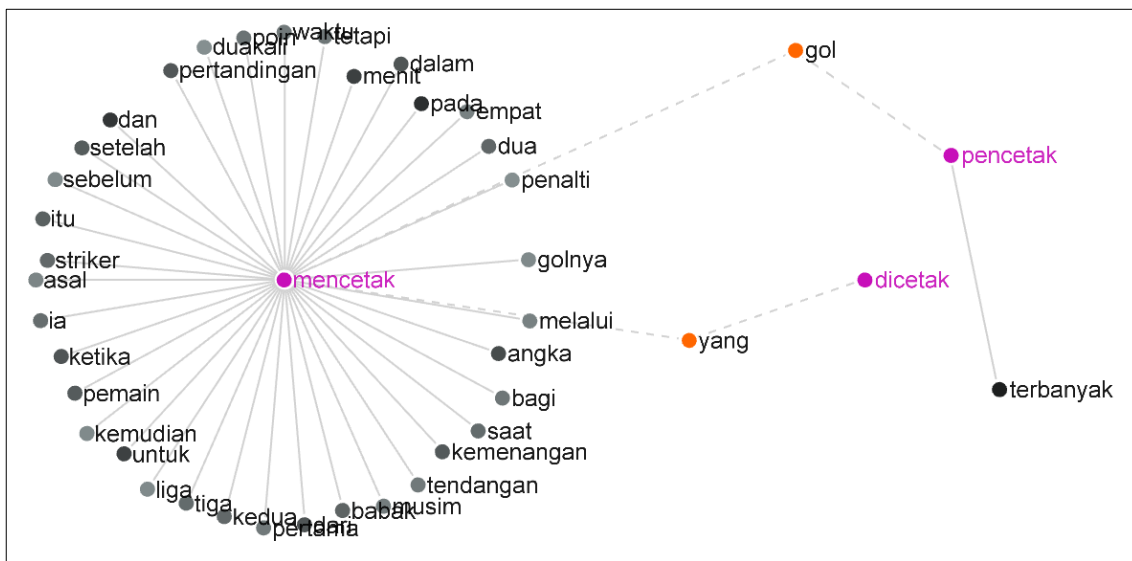


Figure 11: Collocation network of three words from the same headword in the BPPT-PAN corpus in *LancsBox* (LR=5, Stat=LL, Threshold=50.0, View=Type, Filter=None)

Smart search is a *LancsBox* feature, but for Indonesian it is still under development. I am developing intelligent search grammars to facilitate rapid pattern retrieval. By entering PASSIVE, for instance, all verbs with passive voice prefixes such as *di-* and *ter-* will be included in the search results, as indicated by the KWIC lines below (Figure 12). Unlike POS and lemma annotation, however, this feature is not yet fully developed.

	Node	
laksanaan pernyataan bersama yang	dikeluarkan	19 September 2005, kata utusan pemer
n. Meskipun dilakukan penumpasan,	terdapat	190 kematian berkaitan dengan geng ob
but. Pabrik briket PTBA Tanjung Enim	didirikan	1997 dengan kapasiti produksi sekitar
tal investasi 2,8 juta RMB dan modal	dicatatkan	2 juta RMB. Menurut Herman, saham pe
untuk satu bulan hingga 1 tahun tetap	dipatok	2,50 persen. Badan Pengawas Pasar M
lasi tahun kalender Januari-Juli 2007	tercatat	2,81 persen, dan inflasi year on year
si itu setelah kecolongan dua gol dan	tertinggal	2-3 dari tim divisi tiga itu sedangkan
dengan rencana kunjungan itu, yang	dijadwalkan	20 sampai 31 Maret. Saya ingin mendap

Figure 12: Smart search KWIC outcome for PASSIVE in Indonesian

4.3.2. *CQPweb* v.3.3.11

As *CQPweb* is an online corpus query system, users do not need to insert all the required language resource files. In the *CQPweb* server version 3.3.11, the Indonesian TT parameter file and other required resources have been incorporated. Numerous large Indonesian corpora on *CQPweb* have been annotated with POS and headwords (such as LCC Indonesian 2023), so users can immediately begin searching without having to perform any additional steps.

Note that users can also install their own corpus in *CQPweb*. First, navigate to Manage your files under Your Files and Corpora. Then upload your files to the *CQPweb* server from your desktop computer. Next, select Install corpus with TT (Indonesian) under Install your own corpus (Figure 13). There is a one-million token limit for those who wish to utilise language support. If your corpus exceeds one million tokens, *CQPweb* will automatically remove any excess tokens.



Figure 13: The Indonesian TT

Once your corpus is uploaded, you can search it using POS tags and headwords. The list of POS tags can be viewed under the Corpus Info section of the CS UI POS tagset. With the annotated corpus, all *CQPweb* functionalities are now active. For example, tags can be used to filter collocates. Figure 14 shows noun collocates of the node *sepakbola* ‘football’. We can see that some collocates are institutional, such as *federasi* ‘federation’ and *konfederasi* ‘confederation’, and some are proper names, namely, FIFA, UEFA, and FA.

No.	Word	Dice coefficient
1	Federasi	0.07330
2	FIFA	0.03650
3	Asosiasi	0.03430
4	UEFA	0.02830
5	FA	0.02700
6	Konfederasi	0.02480
7	pecinta	0.02040
8	klub	0.02020
9	pemerhati	0.01710
10	badan	0.01530

Figure 14: Top-10 collocates of *sepakbola* in LCC Indonesian 2023 *CQPweb*

We can also retrieve complex patterns by combining headword and POS tags in a query. For instance, the query {cetak} gol _JJ retrieves the following word combinations: all words whose headword is {cetak} followed by *gol* ‘goal’ and ending with any adjective (Figure 15).

{cetak} gol _JJ

Query mode: Simple query (ignore case) [Simple query language syntax](#)

Number of hits per page: 50

Match strategy: Standard

Restriction: None (search whole corpus)

Start query
Reset query

Figure 15: Sample of a query that includes headword and POS based annotation (LR=3, Stat=Dice Coefficient, Threshold=Default, View=Word, Filter=noun)

This query combines orthographic, headword, and POS tags searches. Table 11 provides the user with a number of instances in the concordance lines, such as *pencetak gol*

terbanyak ‘top goal scorers’, *mencetak gol cepat* ‘scored a quick goal’, and *mencetak gol penting* ‘scored an important goal’, among others.

Left Context	Node	Right Context
1 <i>jajaran_NN</i> line-up_SC ‘groups	<i>pencetak_NN gol_NN terbanyak_JJ</i> scorer_NN goal_NN most_JJ of top goal-scorers	<i>Barcelona_NN</i> Barcelona_NN in Barcelona’
2 <i>berhasil_VB</i> succeed_VB ‘managed to	<i>mencetak_VB gol_NN cepat_JJ</i> score_VB goal_NN quick_JJ score a quick goal	<i>pada_IN</i> at_IN at the’
3 <i>juga_RB</i> also_RB ‘also	<i>mencetak_NN gol_NN penting_JJ</i> score_VB goal_NN beautiful_JJ scored a beautiful goal	<i>untuk_SC</i> for_SC for’

Table 11: Sample of concordance lines from a complex query (three randomly picked lines)

5. SUMMARY AND CONCLUSIONS

All the objectives of the present research have been achieved. The first objective was the creation of the Indonesian TT. The creation of an Indonesian TT parameter file was shown in Section 4.1. This file can be used to perform headword and POS annotation on Indonesian corpora using TT. The second objective, the performance of the Indonesian TT, was evaluated in Section 4.2. The Indonesian TT achieves a POS tagging accuracy of 96 per cent on the testbed corpus and between 92 per cent and 96 per cent on alternate testbed corpora. In terms of headword annotation, the Indonesian TT achieves 81 per cent precision on the testbed corpus and between 75 and 81 per cent precision on the alternative testbed corpora. As for the third objective, Section 4.3 showed that support for the Indonesian language was added to both *LancsBox* version 6.0 and *CQPweb*, as of version 3.3.11 onwards.

The fact that the POS tagging accuracy of the Indonesian TT is between 92–96 per cent is the major limitation of the study, since the percentage is slightly lower than other finer-grained tagset, such as the *IPOSTagger*. Despite being 2–4 per cent lower, the current 90+ per cent accuracy standard is widely accepted best practice for POS taggers. Also, the *IPOSTagger* does not output headwords, while the Indonesian TT does. In the future, I hope to address existing issues by revising the resources (expanding the training data and enhancing the quality and quantity of the lexicon), by making the tagset more

granular, revising incorrect headwords and incorrectly labelled entries, and adding multi-word unit entries. Thorough inspections and revision to POS tags and headwords in the resources will be conducted by inter-rater agreement to improve reliability. This is expected to improve also the headword and POS tagging accuracy of the revised Indonesian TT in the future study.

Regardless of the limitations, the availability of Indonesian language support for *LancsBox* and *CQPweb* provides users with more search power. This opens up opportunities to qualitatively or quantitatively amplify data interpretation. Although the project focuses on POS and headword annotation, the findings may also have a bearing on syntactic annotation, as phrase structure can be identified using a combination of POS tags. While the Indonesian TT in this project has been integrated into *LancsBox* and *CQPweb*, nothing prevents users with greater technical expertise from integrating the resources into other systems.

REFERENCES

- Anthony, Lawrence. 2024. *AntConc* v.4.2.4 [Software]. Tokyo, Japan. <https://www.laurenceanthony.net/software/antconc/>
- Bahasa, Pusat. 2005. *Kamus Besar Bahasa Indonesia*. Jakarta: Badan Pengembangan dan Pembinaan Bahasa.
- Bird, Steven, Edward Loper and Ewan Klein. 2019. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly. <https://www.nltk.org/book/>
- Brezina, Vaclav and William Platt. 2024. *#LancsBox X*. <https://lancsbox.lancs.ac.uk/>
- Brezina, Vaclav, Pierre Weill-Tessier and Tony McEnery. 2018. *#LancsBox v.4.x* [Software]. Lancaster. <http://corpora.lancs.ac.uk/lancsbox>
- Brezina, Vaclav, Pierre Weill-Tessier and Tony McEnery. 2020. *#LancsBox v.5.x* [Software]. Lancaster. <http://corpora.lancs.ac.uk/lancsbox>
- Can, Burcu, Ahmet Üstün and Murathan Kurfalı. 2017. Turkish PoS tagging by reducing sparsity with morpheme tags in small datasets. *arXiv*. <https://doi.org/10.48550/arXiv.1703.03200>
- Chung, Siaw-Fong and Meng-Hsien Shih. 2019. *An Annotated News Corpus of Malaysian Malay*. <https://doi.org/10.15026/94451>
- Cohn, Abigail C. and Maya Ravindranath. 2014. Local languages in Indonesia: Language maintenance or language shift? *Linguistik Indonesia* 32/2: 131–148.
- Davies, Mark. 2024. *English Corpora*. [Corpora]. <https://www.english-corpora.org>
- Denistia, Karlina. 2023. Databases on the Indonesian prefixes PE- and PEN. *Journal of Language and Literature* 23/1: 13–24.
- Dinakaramani, Arawinda, Fam Rashel, Andry Luthfi and Ruli Manurung. 2014. Designing an Indonesian part of speech tagset and a manually tagged Indonesian corpus. *Proceedings of the International Conference on Asian Language*

- Processing*. Kuching: Institute of Electrical and Electronic Engineers, 66–69. <https://doi.org/10.1109/IALP.2014.6973519>.
- Eberhard, David M., Gary Francis Simons and Charles D. Fennig eds. 2022. *Ethnologue: Languages of Asia*. Dallas: SIL International.
- Fu, Sihui, Nankai Lin, Gangqin Zhu and Shengyi Jiang. 2018. Towards Indonesian part-of-speech tagging: Corpus and models. http://lrec-conf.org/workshops/lrec2018/W34/pdf/3_W34.pdf
- Gomide, Andressa. 2020. *Corpus Linguistics Software: Understanding Their Usages and Delivering Two New Tools*. Lancaster: Lancaster University dissertation.
- Hardie, Andrew. 2012. CQPweb — Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17/3: 380–409.
- Hardie, Andrew. 2023. *CQPWeb Lancaster*. Lancaster. <https://cqpweb.lancs.ac.uk/>
- 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.
- Lanin, Ivan, Romi Hardiyanto and Arthur Purnama. 2019. *Kateglo Dataset v1.00.20131128*. https://datahub.io/aps2201/kateglo_scrape#resource-kateglo_scrape_zip
- Larasati, Septina Dian, Vladislav Kuboň and Daniel Zeman. 2011. Indonesian morphology tool (MorphInd): Towards an Indonesian corpus. In Cerstin Mahlow and Michael Piotrowski eds. *Systems and Frameworks for Computational Morphology*. Berlin: Springer, 119–129.
- Librian, Andi. 2016. *Sastrawi* [Software]. <https://github.com/sastrawi/sastrawi>.
- Lun, Wong Wei, Mazura Mastura Muhammad, Warid Mihat, Muhammad Syafiq Ya Shak, Mairas Abdul Rahman and Prihantoro Prihantoro. 2023. Vocabulary index as a sustainable resource for teaching extended writing in the post-pandemic era. *World Journal of English Language* 13/3: 181. <https://doi.org/10.5430/wjel.v13n3p181>.
- Maulana, Aditya and Ade Romadhony. 2021. Domain adaptation for part-of-speech tagging of Indonesian text using affix information. *Procedia Computer Science* 179: 640–647.
- Park, Youngmin and Jungyun Seo. 2015. Joint model of Korean part-of-speech tagging and dependency parsing with partial tagged corpus. *International Journal of Knowledge Engineering-IACSIT* 1/1: 49–53.
- Pisceldo, Femphy, Rahmad Mahendra, Ruli Manurung and I Wayan Arka. 2008. A two-level morphological analyser for the Indonesian language. In Nicola Stokes and David Powers eds. *Proceedings of the Australasian Language Technology Association Workshop2008*. 142–50. Hobart: Australian Language Technology Association, 142–150.
- Prihantoro, Prihantoro. 2021a. *An Automatic Morphological Analysis System for Indonesian*. Lancaster: Lancaster University dissertation.
- Prihantoro, Prihantoro. 2021b. An evaluation of MorphInd's morphological annotation scheme for Indonesian. *Corpora* 16/2: 287–299.
- Prihantoro, Prihantoro 2022a. *Buku Referensi Pengantar Linguistik Korpus: Lensa Digital Data Bahasa*. Semarang: Undip Press.
- Prihantoro, Prihantoro. 2022b. SANTI-Morf dictionaries. *Lexicography* 9/2: 175–193.
- Quasthoff, Uwe, Dirk Goldhahn and Thomas Eckart. 2014. Building large resources for text mining: The Leipzig corpora collection. In Chris Biemann and Alexander Mehler eds. *Theory and Applications of Natural Language Processing*. Cham: Springer, 3–24.

- Rashel, Fam. 2016. *Manually Tagged Indonesian Corpus Data*. GitHub. <https://github.com/famrashel/idn-tagged-corpus/tree/a0c7a7409a31f2e6a3103778f2621d222525c450>
- Rashel, Fam, Andry Luthfi, Arawindaamani and Ruli Manurung. 2014. Building an Indonesian rule-based part-of-speech tagger. *Proceedings of the International Conference on Asian Language Processing*. Kuching: Institute of Electrical and Electronic Engineers, 70–73.
- Schmid, Helmut. 1999. Improvements in part-of-speech tagging with an application to German. In Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann and David Yarowsky eds. *Natural Language Processing Using Very Large Corpora*. Text, Speech and Language Technology. Dordrecht: Springer, 13–25.
- Schmid, Helmut. 2024. *TreeTagger: A POS Tagger for Many Languages* [Software]. <https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>
- Scott, Mike. 2024. *WordSmith v.9.0*. Stroud: Lexical Software Analysis. <https://www.lexically.net/wordsmith/>
- Thavareesan, Sajeetha and Sinnathamby Mahesan. 2020. Word embedding-based part of speech tagging in Tamil texts. *Proceedings of the International Conference on Industrial and Information Systems*. Rupnagar: Institute of Electrical and Electronic Engineers, 478–482
- Vasiliev, Yuli. 2020. *Natural Language Processing with Python and spaCy: A Practical Introduction*. San Francisco: No Starch Press.
- Voutilainen, Atro. 1999. A short history of tagging. Hans Van Halteren ed. *Syntactic Wordclass Tagging*. Dordrecht: Springer, 9–21.
- Wicaksono, Alfian Farizki and Ayu Purwarianti. 2010. HMM based part-of-speech tagger for Bahasa Indonesia. *Proceedings of 4th International Malay and Indonesian Language Workshop Jakarta*: Computer Science.

Corresponding author

Prihantoro

Universitas Diponegoro

Faculty of Humanities

Tembalang Campus, Kota Semarang

Jawa Tengah 50275

Indonesia

E-mail: prihantoro@live.undip.ac.id

received: June 2024

accepted: September 2024

The *Multi-Feature Tagger of English* (MFTE): Rationale, description and evaluation

Elen Le Foll^a – Muhammad Shakir^b
University of Cologne^a / Germany
University of Münster^b / Germany

Abstract – The *Multi-Feature Tagger of English* (MFTE) provides a transparent and easily adaptable open-source tool for multivariable analyses of English corpora. Designed to contribute to the greater reproducibility, transparency, and accessibility of multivariable corpus studies, it comes with a simple GUI and is available both as a richly annotated *Python* script and as an executable file. In this article, we detail its features and how they are operationalised. The default tagset comprises 74 lexico-grammatical features, ranging from attributive adjectives and progressives to tag questions and emoticons. An optional extended tagset covers more than 70 additional features, including many semantic features, such as human nouns and verbs of causation. We evaluate the accuracy of the MFTE on a sample of 60 texts from the BNC2014 and COCA, and report precision and recall metrics for all the features of the simple tagset. We outline how that the use of a well-documented, open-source tool can contribute to improving the reproducibility and replicability of multivariable studies of English.

Keywords – software; multivariable analysis; multivariate analysis; open source; corpus linguistics; corpus tool; multi-dimensional analysis; *Python*

1. INTRODUCTION

The addition of multivariable¹ analysis methods to the linguist's toolbox has proven indispensable to shed light on the intricate interplay between diverse linguistic features and the situational/contextual factors that shape them. One of the earliest such methods is multi-dimensional (MD) analysis—a framework pioneered by Douglas Biber in the 1980s and first applied to the study of register variation (see Biber 1984; 1988). What MD analysis and most other multivariable methods used in linguistics have in common

¹ In quantitative linguistics, the terms 'multivariable' and 'multivariate' are frequently equated. Statisticians, however, differentiate between 'multivariable' methods in which several independent variables (predictors) are used to explain or predict a single outcome variable and 'multivariate' ones in which there are two or more dependent (or outcome) variables (Hidalgo and Goodman 2013). As all multivariate methods are, by definition, also multivariable, we use the term 'multivariable' throughout this article as an overarching term encompassing methods such as multivariable linear and logistic regression, factor analysis, cluster analysis, and machine learning methods.



is that they rely on software capable of automatically identifying a large number of linguistic features in many texts. Since the co-occurrences of these linguistic features need to be identified and counted across hundreds or thousands of texts for such analyses to be feasible,² automatic feature taggers are needed. Indeed, they can be said to constitute the backbone of such analysis frameworks. In the context of MD analysis, Biber (2019: 14) stresses that “[a]lthough its importance is not widely recognized, the computer program used for grammatical tagging provides the foundation for MD studies.” This failure to recognise the crucial importance of the tools underlying most quantitative corpus-linguistic analyses is detrimental to the reproducibility of quantitative corpus-linguistic studies. Even when the corpus data are publicly available, if the tools used to process the data are not freely accessible, the results cannot be independently verified. It is also very difficult to test their robustness and generalisability on new data.

This is particularly problematic in the context of the ‘replication crisis’ (first named as such in Pashler and Wagenmakers 2012), which having been first exposed in social psychology has, over the past decade, spared almost no discipline, highlighting the pressing need to improve both the reproducibility of scientific studies and the replicability of their findings.³ Among the many causes of the replication crisis/crises, the lack of published (or otherwise freely accessible) research data, code and software ranks high. It undoubtedly constitutes a major barrier to both the reproducibility and replicability of published research (see, e.g. John *et al.* 2012; Baker 2016; Gewin 2016). (Corpus) linguists, too, are becoming increasingly aware of the implications of non-reproducible methods, as evidenced, for example, by a special issue of *Linguistics* devoted to the replication crisis (Sönning and Werner 2021) and several recent articles and monographs that tackle the issue head-on (e.g., Porte and McManus 2018; Wallis 2021; In’nami *et al.* 2022; McManus 2024).

² Synthesising data from 161 MD studies, Goulart and Wood (2021: 119) report a mean size of corpora used in MD analyses of 5.5 million words (with a very large range from under 10,000 words to over 206 million words). In a meta-analysis of 23 MD studies, Egbert and Staples (2019: 132) report that the mean number of variables entered in MD analyses is around 60.

³ Although the terms are sometimes used interchangeably, we use the term ‘reproducibility’ to refer to the ability to obtain the same results as an original study using the authors’ data and code, whilst we understand ‘replicability’ to be about obtaining compatible results with (more or less) the same method(s) but different data (see, e.g., Berez-Kroeker *et al.* 2018).

2. LITERATURE REVIEW

In the present paper, we introduce a substantially revised and extended version of the *Multi-Feature Tagger of English Perl*: the *MFTE Python* (hereafter MFTE), designed to facilitate the reproducible, multivariable analysis of large English corpora. Although our tool can be used for any type of quantitative corpus analysis, we developed the MFTE primarily with multivariable analyses in mind as it is often possible to manually examine the accuracy of a tagger for one or just a few feature(s). Such a procedure, however, becomes entirely unfeasible when a large(r) number of linguistic variables are entered in an analysis. The MFTE was designed to facilitate large-scale multivariable analyses such as MD studies by aiming to largely eliminate the need for the manual correction of feature identification (‘fix-tagging’, see Section 2.1.3)

In the following, we explain our rationale for the development of a new open-source tagger written in a modern, object-oriented programming language. The rationale in Section 2.1 is followed by the specifications of the tagger in Section 2.2. In Section 2.3, we list the linguistic features it tags and counts and explain how they are operationalised in Section 2.4. In Section 2.5, we explain the tagger’s usage and outputs and then report on a detailed evaluation of the tagger’s output, before discussing its strengths and limitations and concluding with an outlook on future possible uses and developments.

2.1. Rationale for a new multi-feature tagger of English

Several factors motivated the development of the MFTE. They are best summarised as concerns about the reproducibility, transparency, and accessibility of the corpus tools currently most frequently used in multivariable analyses of English. It should be mentioned that these same reasons had originally motivated the development of the *MFTE Perl*, an earlier but considerably less powerful version of the MFTE.⁴ Its documentation (Le Foll 2021) outlined the methodological decisions involved in the selection of its features, details of their operationalisations, and the rationale behind the use of different normalisation units for the feature frequencies. However, no research article about the *MFTE Perl* was submitted as it quickly became apparent that an entirely reworked version of it, ported to *Python* and considerably extended, would fulfil better

⁴ <https://github.com/elenlefol/MultiFeatureTaggerEnglish> (accessed 8 March 2024)

the three-fold objectives of reproducibility, transparency, and accessibility. It is this new *Python* version that we present and evaluate in the present paper.

2.1.1. Reproducibility

Using proprietary software for research constitutes a barrier to reproducibility as only a limited number of researchers (i.e., those who have personal contacts with the developer(s) or the (institutional) means to pay for a licence) can attempt to reproduce the results of published studies. In theory, publishing detailed information about the inner workings of a piece of software is an alternative to making software accessible to all for free and/or publishing its source code. This is essentially what was done in what remains the most cited MD analysis study to date, namely Biber (1988: Appendix II), which includes detailed descriptions of the algorithms of the tagger now widely known as the ‘*Biber tagger*’ (Gray 2019). It was used to identify the features entered in Biber’s (1988) seminal MD analysis. As a result, even though the *Biber tagger* is not available to the wider research community, it is possible to reconstruct it based on this list of algorithms. There is no doubt, however, that such a ‘reconstruction’ is a time-consuming process that requires advanced programming skills, which not all corpus linguists possess.

Demonstrating that reconstruction is possible based on Biber’s (1988: Appendix II) list of algorithms, Nini (2014; 2019) successfully reproduced the functions of the 1988 version of the *Biber tagger* with only very minor differences. The resulting corpus tool, labelled the ‘*Multidimensional Analysis Tagger*’ (MAT), was originally released as freeware in 2013 and subsequently made available under an open-source licence on GitHub in 2020.⁵ MAT allows researchers to conduct reproducible analyses using the linguistic features described in Biber (1988: Appendix II).⁶ More recent MD studies that rely on the *Biber tagger*, however, are not reproducible as the *Biber tagger* has considerably evolved since the 1988 publication (Biber and Egbert 2018: 22; Gray 2019:

⁵ <https://github.com/andrianini/multidimensionalanalysistagger> (accessed 14 December 2023).

⁶ It is worth noting that MAT does more than just tag and count the features used in Biber’s (1988) analysis: it also calculates dimension scores on Biber’s (1988) dimensions of *General Written and Spoken English*, outputs plots with mean values of the tagged texts/corpus against these dimensions, making it ideal for conducting additive MD analyses (Berber Sardinha *et al.* 2019) that rely on Biber’s (1988) *Model of General Spoken and Written English* as a comparison baseline. MAT also assigns each text to one of the eight text types proposed in Biber’s (1989) *Typology of English Texts*. Moreover, the GUI version (for *Windows* only) features a tool for visualising the features of an input text that load on a particular dimension.

46).⁷ Our reproducibility motivation for the development of the MFTE was therefore to allow researchers to conduct reproducible analyses involving a larger number of English lexico-grammatical and semantic features that go beyond those of the 1988-version of the *Biber tagger*, and which users can flexibly and transparently adapt and amend according to their needs.

2.1.2. Transparency

Ensuring that the concrete operationalisations of the features of the MFTE would be as transparent as possible was a further motivating factor for the development of the MFTE. Reproducibility and transparency of research processes are inextricably linked. Not only does the lack of access to the source code of a tagger mean that results cannot be reproduced, but it also means that researchers conducting and evaluating studies that rely on such tools have few means of understanding exactly how the features entered into these analyses were identified. While this is true for many tools used in corpus studies, it is particularly problematic in the context of multivariable studies such as MD analyses, which rely on counting large numbers of linguistic features across many texts, for which manual spot-checking of counts is simply not feasible.

A further aim of the MFTE was therefore to make available both detailed textual explanations of its feature operationalisations, as well as easily accessible source code to be able to examine their concrete operationalisation and, if need be, adapt them to specific language varieties and/or registers, linguistic theories,⁸ and research questions. Whilst it is possible to scrutinise the exact operationalisations of each linguistic feature of MAT, its code structure is relatively complex due to its graphical interface and many additional functions (see footnote 6), which means that only linguists with a strong programming background are likely to be able to edit the source code to introduce customised features/feature operationalisations. In developing the MFTE, we opted for the simplest

⁷ During the second round of revisions for this paper, an anonymous reviewer helpfully pointed out that a new tagger for MD analysis is now available in beta stage. According to the documentation, it is being developed by Kristopher Kyle and colleagues at the Linguistics Department of the University of Oregon in consultation with Douglas Biber's lab in Northern Arizona University: <https://github.com/kristopherkyle/LxGrTgr> (9 September 2024).

⁸ It is worth remembering that POS tagging is fundamentally a method of analysing grammar and morphology. As a consequence, the process inevitably implicitly reflects a specific approach or theory of grammar (McEnery *et al.* 2006; Lindquist 2009: 44–45; Gray 2019: 34)

structure possible: it consists of a single script that can be easily edited by linguists familiar with regular expressions. This brings us to our final, major motivation for developing the *MFTE Python*, which is to contribute to the better accessibility of taggers that can readily be used to conduct multivariable analyses of English.

2.1.3. Accessibility

For some features, both the *Biber tagger* and MAT require (semi-)manual ‘fix-tagging’ procedures to reach high levels of tagging accuracy (Gray 2019: 59–61). In fact, it is recommended that the *Biber tagger* be used in combination with an interactive tag-checker, see Biber and Gray (2013) for details. Although this process allows for the inclusion of linguistic features that cannot reliably be annotated automatically, it requires trained annotators to perform time-consuming manual checks and corrections.⁹ For a tool to be accessible for research projects with little or no funding, we therefore believe that the need for fix-tagging should be reduced to a minimum. Given that most multivariable linguistic methods, including MD analysis, require many different linguistic features to be identified across large corpora, we aimed for high tagging accuracy without the need for human intervention. Moreover, we concluded that the tagger documentation should include a detailed evaluation of the accuracy of the tagging procedure on a representative sample of texts. These evaluation results should be transparently reported for researchers to be able to decide which feature operationalisations are accurate enough for their specific research objectives.

Most standard POS taggers require an additional script to count the number of occurrences of each tag in each text and to normalise these counts if the texts of the corpora are of different lengths. This process adds an extra step in the preprocessing of tagged corpus data for multivariable analyses. In contrast, MAT and the *MFTE Perl* are more accessible in that they output tables of normalised frequencies that can readily be used as input for statistical tools and functions.

Whilst the *MFTE Perl* was designed to be used without fix-tagging, its outputs include normalised frequencies of each feature per text, and its documentation includes

⁹ For instance, for the TOEFL iBT project (with a corpus of 3,839 texts totalling 543,000 words), Gray and Biber (2013: 18) report having recruited and trained ten fix-taggers in addition to two independent coders and a project research assistant for the manual corrections of problematic tags.

detailed results of a thorough evaluation, it nonetheless failed to adequately meet our third aim of accessibility. This is because we believe that accessibility also entails ease of use. Regrettably, the *MFTE Perl* requires a separate installation of the *Stanford POS tagger* (Toutanova *et al.* 2003) which, even with detailed installation instructions, is likely to constitute a barrier for some linguists. This is not the case with the present version of the MFTE (see Section 2.5). Furthermore, the present version of the MFTE was written in *Python*, a more accessible programming language, with a large user base and hence many beginner-friendly tutorials and helper tools (e.g., *Anaconda* and *Anaconda Mini* for installation).

Finally, unlike MAT for *Windows*, the *MFTE Perl* also lacks a graphical user interface (GUI). We consider this to be an important aspect of making open-source tools accessible to the wider research community and argue that a genuinely accessible tagger ought to include a GUI for all major operating systems.

2.2. Specifications of the MFTE

Based on our triple motivation to improve the reproducibility, transparency, and accessibility of multivariable English corpus studies and our survey of the strengths and weaknesses of existing tools in these regards (see Section 2.1), we elaborated the tagger specifications for the MFTE by updating the specifications originally specified for the *MFTE Perl* (Le Foll 2021). These specifications are shown in Table 1.

1. Identify a broad range of lexico-grammatical and semantic features of English
a. that can each be meaningfully interpreted
b. to a satisfactorily high degree of accuracy (with precision and recall rates of > 90%)
c. without the need for human intervention
d. in a broad range of English registers
e. with standard American or British orthography.
2. Output
a. the full tagged texts in plain text format for qualitative analyses of the tagger’s accuracy
b. delimiter-separated values (DSV) files containing both raw and normalised feature counts per text.
3. Be available
a. as source code under a GNU licence for researchers with programming skills to scrutinise, adapt, improve and re-use and
b. as a GUI with adequate documentation for researchers with basic computer skills to be able to run the programme in all major operating systems.

Table 1: Tagger specifications for the MFTE

In what follows, we outline how we set out to meet these specifications. In Section 2.3, we list the linguistic features tagged by the MFTE and, in Section 2.4 we motivate their operationalisations, before explaining how to use the tagger and understand its outputs (Section 2.5).

2.3. Tagset

In line with the MFTE’s intended application for descriptive linguistic analyses (such as multivariable analyses of register variation) as opposed to classification tasks, we explicitly focused on the identification of linguistic features that can be “meaningfully interpretable” (specification criterion 1a). By this, we mean that the “scale and values [of each feature] represent a real-world language phenomenon that can be understood and explained” (Egbert *et al.* 2020: 24).

When designing the feature portfolio of the *MFTE Perl*, Le Foll (2021) examined simplified Hallidayan system networks grammars (see, e.g., Bartlett and O’Grady 2017) in an attempt to minimise researcher bias in the selection of linguistic features. The *MFTE Perl* identifies and counts 75 linguistic features covering lexical density and diversity, fine-grained POS classes, verb tense and aspect, various frequent lexico-grammatical constructions, and a selection of verb semantic categories. The *MFTE Python* builds on this original set of features from the *MFTE Perl*, of which 74 were retained to form the ‘simple tagset’ of the new MFTE. These are listed in Table 2. In addition, the present version of the MFTE features an ‘extended tagset’ with more than 70 additional features, mostly semantic features inspired from Biber *et al.* (1999) and Biber (2006), operationalised on the basis of the features from the simple tagset. A full list of all features including examples and a description of their operationalisation can be found on the MFTE’s [GitHub repository](#). Note that the latest features can be found in the developmental branch of the repository.

Feature	Tag	Example
BE able to	ABLE	<i>It should be able to speak back to you. Would you be able to?</i>
Amplifiers	AMP	<i>I am very tired. They were both thoroughly frightened.</i>
Average word length	AWL	<i>It's a shame that you'd have to pay to get that quality.</i> [AWL = 42/12 = 3.5]
BE as main verb	BEMA	<i>It was nice to just be at home. She's irreplaceable. It's best.</i>
Coordinators	CC	<i>Instead of listening to us, he also told John and Jill but at least his parents don't know yet.</i>
Numbers	CD	<i>That's her number one secret. It happened on 7 February 2019.</i>
Concessive conj.	CONC	<i>Even though the antigens are normally hidden...</i>
Conditional conj.	COND	<i>If I were you... Even if the treatment works...</i>
Verbal contractions	CONT	<i>I don't know. It isn't my problem. You'll have to deal.</i>
Causal conjunctions	CUZ	<i>He was scared because of the costume. Yeah, coz he hated it.</i>
Demonstrative pronouns and articles	DEMO	<i>What are you doing this weekend? Whoever did that should admit it.</i>
Discourse/pragmatic markers	DMA	<i>Well, no they didn't say actually. Okay I guess we'll see.</i>
DO auxiliary	DOAUX	<i>Should take longer than it does. Ah you did. Didn't you?</i>
Determiners	DT	<i>Is that a new top? Are they both Spice Girls? On either side.</i>
Downtoners	DWNT	<i>These tickets were only 45 pounds. It's almost time to go.</i>
Elaborating conjunctions	ELAB	<i>Similarly, you may, for example, write bullet points.</i>
Emoji and emoticons	EMO	 :-(:DD XD :)
Emphatics	EMPH	<i>I do wish I hadn't drunk quite so much. Oh really? I just can't get my head around it.</i>
Existential there	EX	<i>There are students. And there is now a scholarship scheme.</i>
Filled pauses and interjections	FPUH	<i>Oh noooooo, Tiger's furious! Wow! Hey Tom! Er I know.</i>
Frequency references	FREQ	<i>We should always wear a mask. He had found his voice again.</i>
Going to constructions	GTO	<i>I'm not gonna go. You're going to absolutely love it there!</i>
Hedges	HDG	<i>There seemed to be no sort of chance of getting out. She's maybe gonna do it.</i>
HAVE got constructions	HGOT	<i>He's got one. Has she got any?</i>
Hashtags	HST	<i>#AcWri #Buy1Get1Free</i>
Prepositions	IN	<i>The Great Wall of China is the longest wall in the world. There are towers along the wall.</i>
Attributive adjectives	JJAT	<i>I've got a fantastic idea! Cheap, quick and easy fix!</i>
Predicative adjectives	JJPR	<i>That's right. One of the main advantages of being famous...</i>
Lexical density	LDE	<i>It's a shame that you'd have to pay to get that quality.</i> [LDE = 3/14 = 0.21]
Like	LIKE	<i>Sounds like me. And just like his father. I wasn't gonna like do it.</i>
Modal can	MDCA	<i>Can I give him a hint? You cannot. I can't believe it!</i>
Modal could	MDCO	<i>Well, that could be the problem. Could you do it by Friday?</i>
Modals may and might	MDMM	<i>May I have a word with you? But it might not be enough.</i>
Necessity modals	MDNE	<i>I really must go. Shouldn't you be going now? You need not have worried.</i>
Modal would	MDWO	<i>Wouldn't you like to know? I'd like to think it works.</i>
Will and shall modals	MDWS	<i>It won't do. Yes, it will. Shall we see?</i>
Noun compounds	NCOMP	<i>the dungeon entrance; this rare winter phenomenon</i>
Total nouns (including proper nouns)	NN	<i>on Monday 6 Aug, the U.S., on the High Street, comprehension</i>

Table 2: Features of the MFTE's simple tagset

Feature	Tag	Example
@mentions	NN	@gretathunberg @MSF_france
BE-passives	PASS	He must have been burgled . They need to be informed .
Perfect aspect	PEAS	Have you been on a student exchange? He has been told .
GET-passives	PGET	He's gonna get sacked . She'll get me executed .
It pronoun reference	PIT	It fell and broke. I implemented it . Its impact is unproven.
Place references	PLACE	It's not far to go. I'll get it from upstairs . It's downhill .
Politeness markers	POLITE	Can you open the window, please ? Would you mind giving me a hand?
s-genitives	POS	the world's two most populous country; my parents' house
Reference to the speaker/writer and other(s)	PP1P	We were told to deal with it ourselves . It's not ours either.
Reference to the speaker/writer	PP1S	I don't know. It isn't my problem. Nor is it mine .
Reference to addressee(s)	PP2	If your model was good enough, you 'd be able to work it out.
Single, female third person	PP3f	She does tend to keep to herself , doesn't she ?
Single, male third person	PP3m	He is beginning to form his own opinions. I trust him .
Other personal pronouns	PPother	One would hardly suppose that your eye was as steady as ever.
Progressive aspect	PROG	He wasn't paying attention. I'm going to the market.
Quantifiers	QUAN	Such a good time in half an hour. She's got all these ideas. It happens every time.
Quantifying pronouns	QUPR	She said addressing nobody in particular. Somebody will.
Question tags	QUTAG	Do they? Were you? It's just it's repetitive, isn't it?
Other adverbs	RB	Unfortunately , that's the case. Exactly two weeks.
Particles	RP	I'll look it up . It's coming down . When will you come over ?
So	SO	She had spent so many summers there. So , there you go.
Split auxiliaries/infinitives	SPLIT	I would actually drive. You can just so tell. I can 't imagine it.
Stranded prepositions	STPR	We've got more than can be accounted for . Open the door and let them in .
Subordinator <i>that</i> omission	THATD	I mean [THATD] you'll do everything. I thought [THATD] he just meant our side.
<i>That</i> relative clauses	THRC	I'll just run a cable that goes from here to there.
<i>That</i> subordinate clauses	THSC	Did you know that the calendar we use today was started by Julius Caesar?
Time references	TIME	It will soon be possible. Now is the time. I haven't it yet .
Reference to more than one non-interactant and single <i>they</i> reference	TPP3t	The text allows readers to grapple with their own conclusions. Do you see them ?
Lexical diversity	TTR	It's a shame that you'd have to pay to get that quality . [TTR = 12/14 = 0.85]
URL and e-mail addresses	URL	www.faz.net https://twitter.com smith@gmail.au
Past tense	VBD	It fell and broke . I implemented it. If I were rich.
Non-finite verb -ing forms	VBG	He texted me saying no. He just started laughing .
Non-finite -ed verb forms	VBN	These include cancers caused by viruses. Have you read any of the books mentioned there?
Imperatives	VIMP	Let me know! In groups, share your opinion and take notes.
Present tense	VPRT	It's ours . Who doesn't love it? I know .
Direct WH-questions	WHQU	What's happening? Why don't we call the game off? How ?
WH subordinate clauses	WHSC	I'm thinking of someone who is not here today. Do you know whether the banks are open?
Negation	XX0	Why don't you believe me? There is no way. Nor am I.
Yes/no questions	YNQU	Have you thought about giving up? Do you mind ?

Table 2: Features of the MFTE's simple tagset (Continuation)

The MFTE performs feature extraction in several steps. First, each text is tagged with basic POS labels using the POS tagger from the *stanza Python* library (Qi *et al.* 2020). Then, several loops of rule-based algorithms are run to refine some of the analyses of the POS tagger and to identify further linguistic features on the basis of syntactic patterns defined by a combination of regular expressions and dictionary lists. The lists are based on Biber (1988), Biber *et al.* (1999), Biber (2006), and the COBUILD (Sinclair *et al.* 1990).

Whilst many of the features of the simple tagset may look superficially like the 1988 version of the *Biber tagger* (and hence also to the MAT), their operationalisations often differ substantially. In particular, the algorithms used to capture verb tense, aspect, and voice are fundamentally different. For example, rather than tag the perfect aspect onto the auxiliary HAVE as in MAT, the MFTE assigns the ‘perfect aspect tag’ (PEAS) to the past participle form of the verb, whilst the auxiliary is marked for tense with either the VBD or the VPRT tag. Similarly, the ‘passive voice tag’ (PASS) is assigned to the past participle form rather than to the verb BE. These new operationalisations make possible the creation of a distinct, linguistically meaningful, VBN variable, which only includes non-finite uses of past participle forms. As the MFTE also identifies verbs in the progressive aspect (PROG) and the *going-to* construction (GTO), non-finite uses of present participle forms can also be accounted for in a separate variable (VBG). Moreover, these verb feature operationalisations are optimised for a range of syntactic patterns, e.g., by allowing for various combinations of intervening adverbs, negation, and paralinguistic sounds in verb phrases involving auxiliaries. Other lexico-grammatical features from the MFTE’s simple tagset that are not typically found in lexico-grammatical taggers include question tags (QUTAG), imperatives (VIMP), and emojis or emoticons (EMO).

Various strategies were employed to deal with multifunctional/polysemous lexical items: in many cases, they were included in the semantic categories corresponding to their most frequent functions (e.g., *only* is assigned to the category of downtoners rather than hedges). When this proved too error-prone, infrequent items were excluded. Moreover, for two highly frequent multifunctional words (*so* and *like*) separate feature categories were created to capture all potentially problematic cases, i.e., all occurrences of *so* and all occurrences of *like* tagged as a preposition or adjective by the stanza POS tagger. Both words are frequently used as discourse markers and fillers in conversation. These uses are

very difficult to automatically disambiguate. Consequently, if they are not excluded, they run the risk of severely distorting the frequencies of adverbs and prepositions in conversational texts. More recent versions of the *Biber tagger* make use of different algorithms and probabilities depending on the mode/register of the texts to be tagged to (partially) circumvent this issue (Gray 2019: 46). We believe that such a procedure bears the risk of predetermining patterns of occurrences and have therefore opted to relegate these items to separate feature categories for *like* (when not tagged as a verb) and *so*, instead. While this is by no means a perfect solution, it allows users of the MFTE to decide—depending on the resources available to them—whether to perform manual fix-tagging to assign the correct functional tag to each occurrence of *so* and *like*, or to exclude these tokens by not entering the counts of the SO and LIKE variables in their analyses.

When using POS taggers, Brezina (2018: 192) suggests removing paralinguistic sounds (e.g., *um*, *er*, *mhm*, *mm*) from the spoken data as these are frequently misidentified as nouns, thus potentially vastly overestimating the noun count in natural conversation corpora in which these sounds have been transcribed. Rather than remove these tokens as part of the pre-processing of the corpus texts, we opted to retain them as they constitute a defining characteristic of natural spoken language. The MFTE counts these as a self-contained linguistic variable (FPUH), which, like all the linguistic features counted by the MFTE, the user is free to either include or exclude from their analyses, depending on their texts (considering, for instance, how reliably these were transcribed across of the texts under study) and research questions.

A final major departure from the feature extraction principle applied by some lexico-grammatical taggers that is worth mentioning concerns the treatment of multiword items. The 1988 version of the *Biber tagger* and MAT (Nini 2014; 2019), for example, tag the first token of the multiword *on the other hand* as a conjunct and assigns NULL tags to the remaining tokens, as illustrated in (1) below. This means that, when tagged with these taggers, the string *on the other hand* counts as just one conjunct. In contrast, the MFTE assigns the conjunct tag (CONJ) in addition to the usual preposition, determiner, adjective, and noun tags, as shown in (2). NULL tags were only retained for a select few multiword units that largely resist compositional analysis: *of course* (tagged as a discourse marker, DMA), *all right* (DMA) and *no one* (tagged as a quantifying pronoun, QUPR).

- (1) CONJ the_NULL other_NULL hand_NULL
- (2) on_IN CONJ the_DT other_JJ hand_NN

The three semantic verb categories employed in Biber (1988) are only used in the MFTE for the identification of the *that*-omissions in subordinate clauses (THATD). Instead, the MFTE’s extended tagset adopts, with some minor corrections, the semantic categories described in Biber *et al.* (1999) and Biber (2006). As a result, in addition to ten semantic verb features, the extended tagset also includes fine-grained semantic categories for adverbs, nouns, and adjectives, as well as combined grammatico-semantic features such as ‘*to* clauses preceded by verbs of desire’ (ToVDSR) or ‘stance nouns without prepositions’ (NSTNCothers). The extended tagset also includes comparative and superlative constructions (see [GitHub repository](#) for details).

The user should be aware that the tables of counts generated when using the extended tagset include some composite features. These consist of aggregates of individual feature counts, e.g., the variable PASSall contains the sum of the counts of *BE*-passives (PASS) and *GET*-passives (PGET), two features from the simple tagset. It goes without saying that, to avoid redundant correlations, these aggregate counts should not be entered in MD analyses together with their respective individual feature counts. In proposing this large number of lexical, grammatical, and semantic features, we are not suggesting that it makes sense to enter them all in multi-feature analyses of English, but rather that, with the help of our detailed documentation, users can make informed, linguistically motivated choices as to which features are relevant and reliable enough (see evaluation results in Section 4.3) for their use cases.

2.4. Usage

Since the alpha version was first released in April 2023, the MFTE has been accessible under an open-source GPL-3.0 licence. It can be downloaded from a dedicated GitHub repository at <https://github.com/mshakirDr/MFTE>. Detailed installation and usage instructions can be found on the main page of the [GitHub repository](#).

The MFTE relies on commonly utilised *Python* dependencies whose installation are not particularly involved. To expand its accessibility to non-programming linguists, we have also released the programme as an executable file for *Windows*. Once this executable file has been downloaded, the GUI can be run without installing *Python* or any

dependencies and is therefore the most user-friendly version of the MFTE (currently only for *Windows*; see Figure 1).

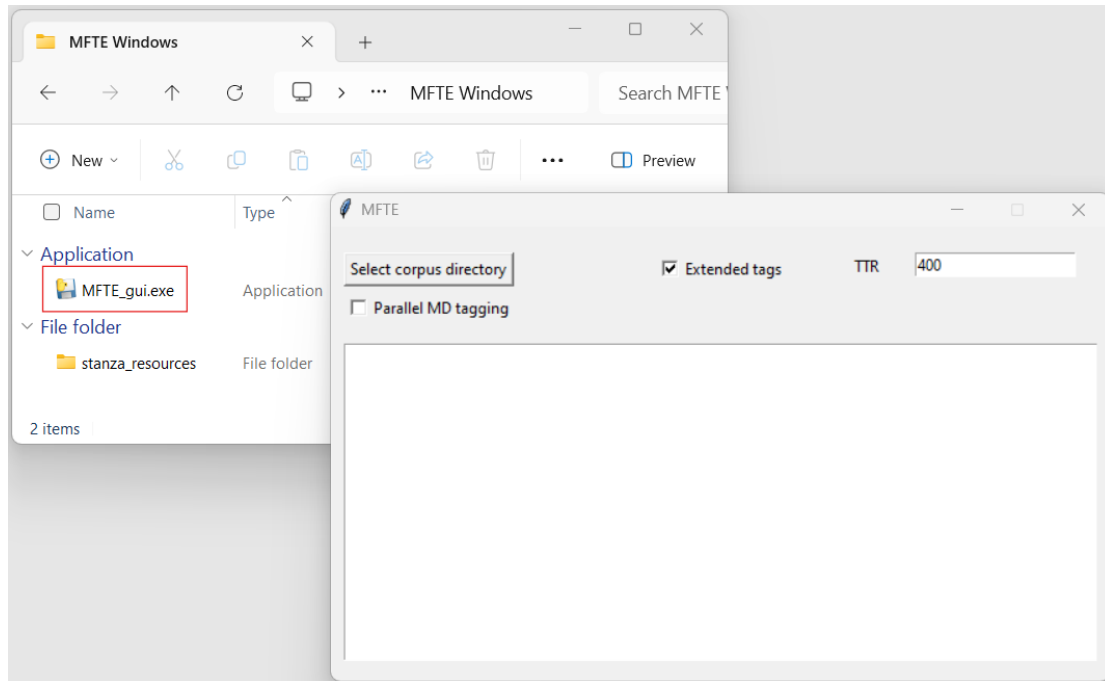
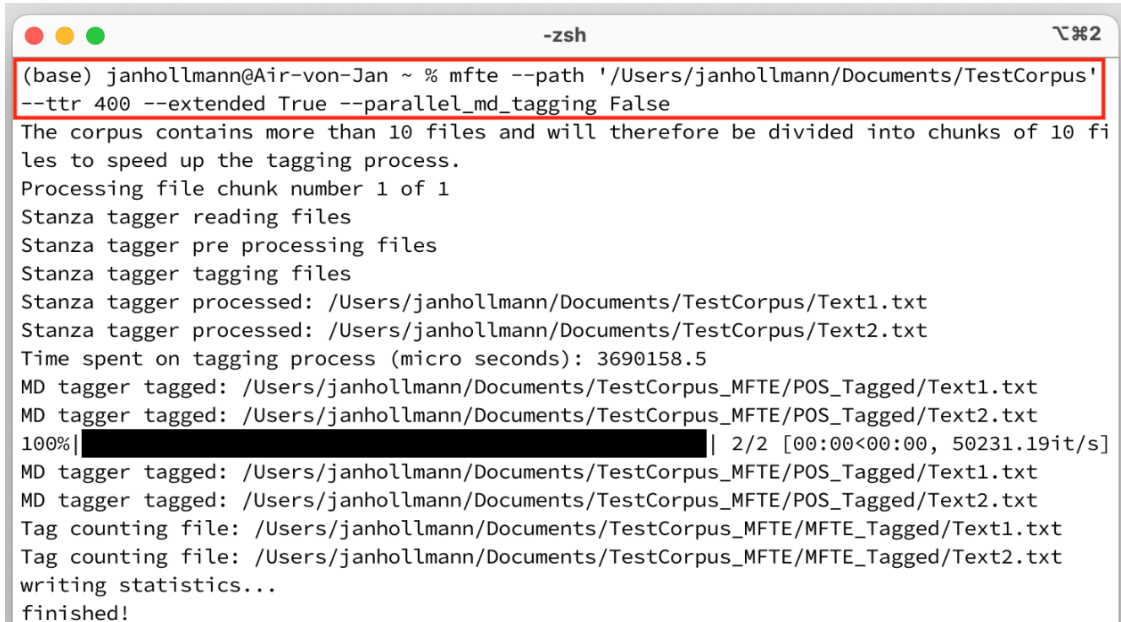


Figure1: Screenshot of the downloaded MFTE GUI executable for *Windows* (MFTE_gui.exe), alongside the running MFTE GUI application

For *Mac OS*, *Linux*, and *Windows*, both the command-line version and the GUI version can be run from a terminal. The command-line version is straightforward (see Figure 2), requiring only four arguments:

1. the path to the folder containing the text files to be tagged;
2. the number of words used to calculate the type-token ratio (TTR)—as in the MAT, the default is 400 but this should be set lower if any texts are shorter than 400 words;
3. an option to use the extended tagset (the default is True);
4. an option to tag multiple files in parallel (this reduces running time for large corpora significantly but may also cause high CPU usage, hence the default setting is False).



```

(base) janhollmann@Air-von-Jan ~ % mfte --path '/Users/janhollmann/Documents/TestCorpus'
--ttr 400 --extended True --parallel_md_tagging False
The corpus contains more than 10 files and will therefore be divided into chunks of 10 fi
les to speed up the tagging process.
Processing file chunk number 1 of 1
Stanza tagger reading files
Stanza tagger pre processing files
Stanza tagger tagging files
Stanza tagger processed: /Users/janhollmann/Documents/TestCorpus/Text1.txt
Stanza tagger processed: /Users/janhollmann/Documents/TestCorpus/Text2.txt
Time spent on tagging process (micro seconds): 3690158.5
MD tagger tagged: /Users/janhollmann/Documents/TestCorpus_MFTE/POS_Tagged/Text1.txt
MD tagger tagged: /Users/janhollmann/Documents/TestCorpus_MFTE/POS_Tagged/Text2.txt
100%|████████████████████████████████████████████████████████████████████████████████| 2/2 [00:00<00:00, 50231.19it/s]
MD tagger tagged: /Users/janhollmann/Documents/TestCorpus_MFTE/POS_Tagged/Text1.txt
MD tagger tagged: /Users/janhollmann/Documents/TestCorpus_MFTE/POS_Tagged/Text2.txt
Tag counting file: /Users/janhollmann/Documents/TestCorpus_MFTE/MFTE_Tagged/Text1.txt
Tag counting file: /Users/janhollmann/Documents/TestCorpus_MFTE/MFTE_Tagged/Text2.txt
writing statistics...
finished!

```

Figure 2: Screenshot of the command line version of the MFTE (example command highlighted in red)

2.5. Outputs

The outputs of the MFTE are “recorded and encoded in the annotated corpus” such that they are “explicit and recoverable” (McEnery *et al.* 2006: 31). To this end, the MFTE creates two versions of each processed text files:

1. the texts as tagged using the underlying POS tagger (in a subfolder labelled ‘POS_Tagged’) and
2. the texts as subsequently tagged by the MFTE (in a subfolder labelled ‘MFTE_Tagged’).

This ensures that it is possible to trace back the origin of tagging errors and to rectify them by either modifying the code of the tagger or manually fix-tagging the tagger output.

The MFTE also outputs three tables of feature counts as comma-separated text files. The filenames include the name of the folder in which the original texts are stored as a prefix. The folders are named following this pattern:

1. [prefix]_normed_100words_counts.csv
2. [prefix]_normed_complex_counts.csv
3. [prefix]_raw_counts.csv

Each CSV file consists of a data matrix in which each row corresponds to a text file from the tagged corpus and each column to a linguistic feature. In all three tables, the first five columns of the count tables are identical: the first column lists the filenames, the second the total number of words in each text as used for word-based normalisation (excluding fillers, see online appendix). The remaining columns correspond to the linguistic features listed in Table 2 for the simple tagset followed by those of the extended tagset (see online appendix), if this option was selected.

The difference between the three tables that the MFTE outputs is that the first contains feature frequencies normalised per 100 words. In the second table, ‘complex’ normalised feature frequencies are calculated based on the three normalisation baselines listed in the fifth column of the table in the appendix. This means that, in this table, the present tense variable (VPRT) represents the percentage of finite verbs in the present tense. As such, it can range from zero, i.e., texts in which no single verb is in the present tense to 100, i.e., texts that are exclusively in the present tense, and therefore does not violate the assumptions of the binomial distribution required for many statistical methods commonly used in corpus linguistics research (Wallis 2020: 56). For details on the normalisation units used for each feature and the rationale behind these choices, see Le Foll (2021: 20–23; 2024: 120–124). Lastly, the MFTE outputs a table of raw counts. Unless the text samples of the examined corpus are all the same length, this table should not be used *as is* for any statistical analyses; however, it may be used by researchers who wish to implement their own normalisation baseline(s). We have also found it very useful for test purposes, i.e., to check how the tagger deals with certain strings.

3. EVALUATING TAGGER ACCURACY

Given that taggers provide “the foundation for MD studies” (Biber 2019: 14) and other multivariable studies, their accuracy is crucial for drawing reliable conclusions that can contribute to building cumulative knowledge. In a comparison of the accuracy of just four linguistic features as identified by three different taggers, Picoral *et al.* (2021) reported differences large enough to lead to significantly different conclusions. Despite this, Goulart and Wood (2021: 123) concluded that tagger accuracy is underreported in the majority of published MD studies. This is not surprising, as the large number of linguistic features typically entered in such analyses makes comprehensive evaluations of tagger

accuracy an arduous task. In the context of most small- to mid-scale projects, it is simply not feasible. This is why we believe that, to meet our aim for providing the research community with an accessible tagger, it is necessary to conduct and publish the results of a comprehensive evaluation of the accuracy of the MFTE.

When evaluating the performance of a tagger, several accuracy measures can be used, as shown in Table 3 below. On the one hand, the number of correct tags, i.e., true positives, can be counted and, on the other hand, the number of incorrect tags, i.e., false positives. The simplest measure of accuracy is ‘precision’: the ratio of true positives to all tags assigned by the tagger. ‘Recall’, by contrast, is the ratio of true positives to all instances of a given tag in the data. It takes into account both true positives and false negatives (instances where a particular tag should have been assigned by the tagger but was not). In other words, recall provides a measure of the tagger's ability to correctly identify and classify all instances of a particular feature. While both precision and recall provide valuable information about a tagger's performance, they only give a partial picture of its accuracy. If precision is high but recall is low, the tagger may be too conservative, tagging only those instances about which it is particularly confident. Conversely, a tagger with low precision but high recall will assign some tags too liberally.

Term	Definition
True positive	Feature correctly identified as X
True negative	Feature correctly identified as not X
False positive	Feature incorrectly identified as X
False negative	Feature incorrectly identified as not X
Precision	$\text{True positive count} / (\text{true positive count} + \text{false positive count})$
Recall	$\text{True positive count} / (\text{true positive count} + \text{false negative count})$
F1 score	$2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$

Table 3: Summary of the terminology frequently used in tagger performance evaluations

In an ideal world, a tagger would have 100 per cent precision (i.e., all assigned tags are correct) and 100 per cent recall (i.e., all features are labelled with all the correct tags). In practice, however, attempts to increase precision on any one feature will usually result in lower recall rate for that particular feature (and many others) and vice versa. An important aspect of tagger development therefore consists in striking the appropriate balance between precision and recall. If it is feasible to complement automatic tagging with a manual fix-tagging phase, then it makes sense to prioritise recall. If, however, a tagger is to be used without any manual intervention, both precision and recall are important. This

is why a third accuracy measure is often calculated: the F1 score, which combines precision and recall (see Table 3). It ranges from 0 (entirely inaccurate) to 1 (perfectly accurate) and provides a single number that can be used to compare the performance of different taggers.

A common strategy in POS tagger evaluations is to report one overall measure of per-token accuracy across all tags (typically slightly over 97%, e.g., Manning 2011). This is common practice in NLP but invariably leads to misleadingly inflated results. Indeed, some of the most frequent tokens, in particular punctuation markers and determiners, are both highly frequent and extremely easy to ‘get right’ (e.g., *the* and *I*). By contrast, per-sentence accuracy rates of POS taggers tend to be considerably more modest (hovering around 50–57%) and considerably lower rates for non-standard varieties and registers for which there is little training data (Manning 2011).

Although it is crucial for users of a tagger to be aware of its per-feature accuracy to gauge the reliability of its annotation, few taggers have documentation that includes detailed results of a comprehensive accuracy evaluation study. The *Biber tagger* constitutes an exception. We are aware of two comprehensive evaluations of its tagging accuracy: Biber and Gray (2013: 16–18) and Gray (2015). The first, however, only reports post-fix-tagging precision and recall rates (Biber and Gray 2013: Appendices C and D). In other words, users of the *Biber tagger* who do not have access to the fix-tagging scripts used as part of this project cannot expect similar accuracy levels when tagging their own texts. Moreover, in Biber and Gray (2013: 18) the same 5 per cent sample of texts that were originally manually checked for tagging errors was analysed to calculate the reported precision and recall rates after fix-tagging. This procedure entails the risk of inflating the tagger’s accuracy metrics as the fix-tagging scripts may have been constructed based on errors specific to the sample. By contrast, Gray (2015) reports both “initial reliability rates” of the uncorrected *Biber tagger* output and “final reliability rates after [fix-tagging] scripts” (Gray 2015: Appendix B) on excerpts of a random sample of 15 research articles across different disciplines and registers. On close inspection of the initial reliability rates, it transpires that not all of the features of the *Biber tagger* (or rather the version used in this particular project) seem suitable for use without some manual or semi-automatic ‘fix-tagging’ procedure, as recall and precision rates for some features are well below 90 per cent.

An anonymous reviewer drew our attention to the fact that several recent PhD projects conducted at Northern Arizona University (the ‘home’ of the *Biber tagger*) include small-scale evaluations of the accuracy of more recent versions of the *Biber tagger* for selected features. Goulart (2022: Appendix A) reports recall and precision rates for 29 features in L1 and L2 academic writing. Wood (2023: 59–63) does so for 28 features in a sample of statutory texts, while Dixon (2022: 71–72) evaluated the tagger’s accuracy for 20 features in 17 texts representing gaming language. Before fix-tagging, Goulart (2022: Appendix A) reports precision and recall rates well above 90 per cent for all the selected features except nominalisations. In Wood’s (2023) corpus of statutory language, on the other hand, accuracy is much lower for many of the selected features (e.g., noun and adjective complement clauses, agentless passives, clausal and phrasal coordinated conjunctions). These results confirm that the accuracy of taggers varies widely across different linguistic features and depends on the type of texts being tagged. While it is inevitable that such labour-intensive evaluations cannot realistically be carried out on large samples, it is problematic that most of these evaluations do not report the number of tags on which recall and precision were calculated. This means that confidence intervals around the reported accuracy metrics cannot be computed (see Section 4.2). Dixon’s (2022: 72) evaluation constitutes an exception: the figures reveal that some of the reported accuracy metrics were calculated on as few as three or four occurrences, thus confirming that the results of such small-scale, project-specific evaluations cannot meaningfully be used in other projects.

The documentation of the *MFTE Perl* (Le Foll 2021) includes a detailed evaluation of the tagger on a stratified random sample from the *British National Corpus 2014* (BNC2014; Brezina *et al.* 2021). Samples of 24 texts covering written and spoken British English from a range of registers were tagged using the *MFTE Perl*. The resulting 31,311 tags were manually annotated by two linguists. For each tag, the annotators had three options: 1) mark it as correct, 2) mark it as incorrect and assign the correct tag, or 3) mark it as an unclear or ambiguous context/token for which no tag could reliably be assigned. Most of the 75 features of the *MFTE Perl* achieved a “satisfactorily high degree of accuracy” (Le Foll 2021: 14) with high rates of both recall and precision. F1 scores below 90 per cent were reported for 10 features. These include two rare features for which Le Foll (2021: 45) warns that accuracy rates are unreliable because they are based on very few occurrences.

4. EVALUATION

In the following, we report on the accuracy of the simple tagset of the MFTE on British and American English texts from a range of registers. Given how misleading overall accuracy rates may be (see Section 3), we report recall, precision, and the combined F1 measure for each feature of the simple tagset, thus allowing users to make an informed decision as to which set of features they wish to include in their analyses and which they would like to exclude depending on their research questions and the characteristics of their data. Whenever possible, we also report 95 per cent confidence intervals around these accuracy metrics.

4.1. Data

In line with the specifications of the MFTE (see Section 2.2), for the evaluation of the *MFTE v.1.0*, we chose a stratified random sample of 30 texts from the BNC2014 and 30 texts from the *Corpus of Contemporary American English* (COCA; Davies 1990) representing diverse registers, as shown in Table 4. To maximise the potential for text/topic-specific tagging errors, we analysed samples of roughly 1,000 tokens from the longer texts in our evaluation sample. This was motivated by an observation from the evaluation of the *MFTE Perl* in which most tagging errors were found to be text/topic-specific and therefore highly clustered (see Le Foll 2021: 25–43). The sampled texts from the spoken subcorpus of the BNC2014 were pre-processed to remove the anonymisation tags and metadata following the procedure documented in Le Foll (2021: 28).

Corpus	Subcorpus	Number of texts	Number of tags
BNC2014	Academic writing	3	2,617
BNC2014	E-Language: Blogs	3	2,092
BNC2014	E-Language: E-Mails	2	1,964
BNC2014	E-Language: Forums	2	2,822
BNC2014	E-Language: Reviews	2	3,291
BNC2014	E-Language: Social Media Posts	3	3,263
BNC2014	E-Language: Text Messages	1	396
BNC2014	Fiction	3	4,144
BNC2014	News: Magazines	2	2,363
BNC2014	News: Newspapers	6	5,036
BNC2014	Spoken: Conversation	3	5,312
COCA	Academic writing	3	2,877
COCA	E-Language: Blogs	3	3,337
COCA	E-Language: Web Pages	3	2,416
COCA	Fiction	3	3,129
COCA	News: Magazines	3	2,202
COCA	News: Newspaper Articles	4	2,342
COCA	News: Newspaper Opinion Pieces	4	2,993
COCA	Spoken: Conversation	4	6,015
COCA	Spoken: TV/Movies	3	2,541
Totals		60	61,154

Table 4: Evaluation data

4.2. Methodology

The 60 files tallied in Table 4 were tagged using the *MFTE Python v.1.0* with its simple tagset. The resulting tagged text files were then converted to a spreadsheet format for manual evaluation.¹⁰ Each tag was marked as either correct, unclear, or incorrect. In the case of an incorrect tag, a corrected version of the tag was added to the corresponding column. In addition, tags were added where they were missing.¹¹ These 60 spreadsheet files were subsequently processed and merged using custom *R* functions.

¹⁰ The evaluation was performed by the first author and her research assistant, Tatjana Winter, whom we thank for her meticulous work.

¹¹ For details of the procedure, see Le Foll (2021: 29–33) and Le Foll and Shakir (2023).

Statistics and data visualisations were computed in *R* and *Python* (see [GitHub repository](#) for data and code). Bootstrap simulation was used to calculate 95 per cent confidence intervals (CI) for the precision, recall and F1 measures of each feature based on the results of 1,000 bootstrapped samples (in a procedure inspired by Picoral *et al.* 2021). We performed this task in *Python* and applied it to all tags for which there were more than 100 occurrences in the 60 evaluation files (see Table 4).

4.3. Results

Of the 61,154 manually reviewed tags, 294 (0.48%; 95% CI [0.43–0.54%]) were deemed by the human annotators to be ‘unclear’ due to misspellings, text processing errors, ambiguous contexts, or fragmented sentences (particularly in the COCA data where 5% of each text was removed for copyright reasons). A further 8,506 tags were excluded from the evaluation metrics because they are not tallied in the tables of frequencies generated by the MFTE. As these include all the tags denoting punctuation marks, as well as foreign words, symbols, and other non-word tokens, they would also have considerably inflated the overall accuracy metrics. Excluding these, the number of correctly assigned tags across the 74 linguistic features of the MFTE simple tagset was 51,139. This corresponds to an overall precision of 97.13 per cent (95% CI: 96.99–97.23%) across the 60 evaluation files. Figure 3 shows the recall, precision and F1 measures for each feature of the simple tagset with at least 100 occurrences in the evaluation corpus. The error bars correspond to bootstrapped 95 per cent CI. The colours indicate how many times each tag occurred across the 60 evaluation files (note that the colour scale is logarithmic).

Twelve features from the simple tagset did not meet the precision and/or recall rates of at least 90 per cent stipulated in the tagger specifications (see Section 2.2), although five of these do have F1 scores above 90 per cent. The least accurate features are stranded prepositions (STPR), verbs in the imperative (VIMP), *that* omission (THATD), WH-questions (WHQU), non-finite past participle forms (VBN), and *GET*-passives (PGET).

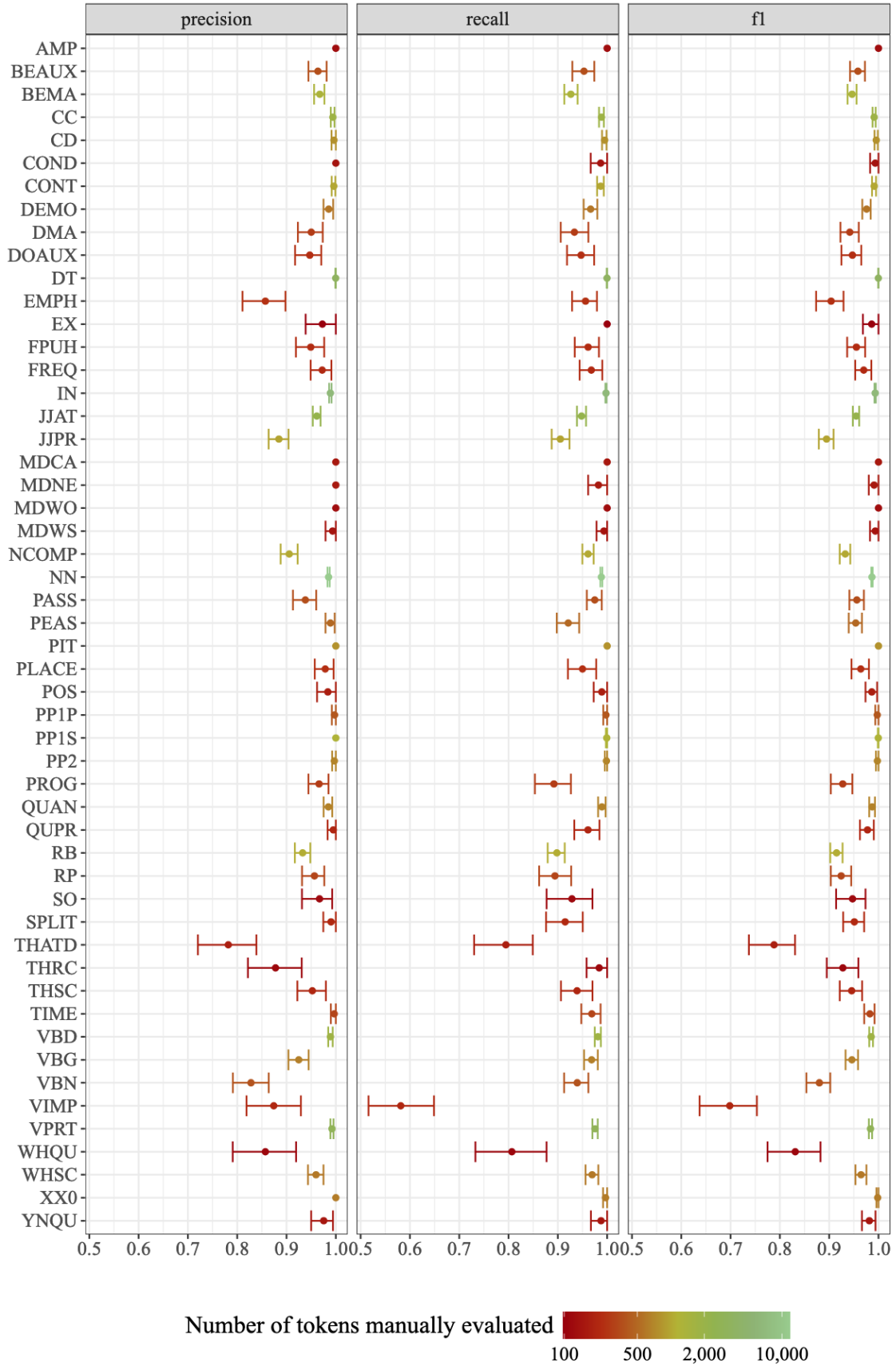


Figure 3: Precision, recall, and F1 score with bootstrapped 95 per cent CI for features with ≥ 100 occurrences across the 60 evaluation files

Figure 4 displays the most frequent tagger errors as clusters of red points. The y-axis lists the tags as assigned by the MFTE, whilst the x-axis corresponds to those assigned by the two human annotators. For the sake of readability, only the most frequent tags are included in Figure 4. The code provided in the repository may be used to compute and examine the full matrix. The figure shows that notable clusters of errors involve the confusion of infinitives (VB) vs. imperatives (VIMP), attributive (JJAT) vs. predicative (JJPR) adjectives, non-finite past participles (VBN) vs. finite verbs in the perfect aspect (PEAS), and WH-questions (WHQU) vs. WH-subordinate clauses (WHSC). We can also see that several clusters of red points involve the tag ‘NONE’. NONE is not part of the tagset. We used this tag as a placeholder to indicate that the MFTE assigned an unwarranted second or third-order tag to this token (e.g., when two adjacent nouns were incorrectly identified as a noun compound by the tagger), or to indicate that a necessary tag was omitted (e.g., if *am* in *I am happy* was assigned the present tense tag, but not the one for *BE* as a main verb).

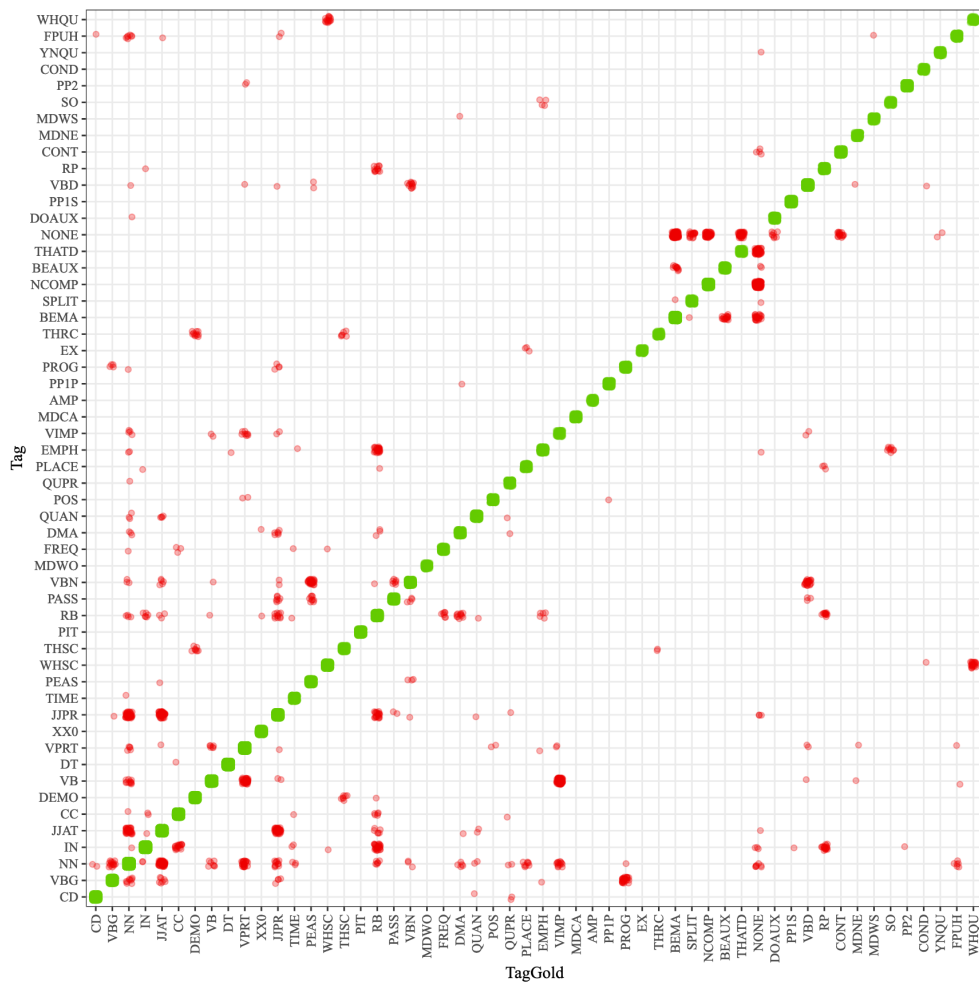


Figure 4: Confusion matrix showing mismatches between the MFTE output (Tag) and the human annotators' assessments (TagGold) in red

As was to be expected, the MFTE performs slightly less well on internet and spontaneously spoken registers than on professionally written and edited texts such as those typically found in academic writing and fiction (see Table 5). For most purposes, however, these differences are unlikely to be significant. The code provided in the repository also allows for the calculation of per-feature accuracy measures for each of the six broad register categories of the BNC2014 and COCA.

Register	Accuracy	Lower CI	Upper CI
Academic writing	97.83	97.37	98.23
E-language	96.56	96.27	96.83
Fiction	97.89	97.49	98.24
News writing	97.77	97.50	98.02
Spoken	96.62	96.28	96.94

Table 5: Overall precision of the MFTE across the broad register categories

5. DISCUSSION

Based on the results of the evaluation, we now examine the extent to which the MFTE meets the tagger specifications elaborated in Section 2.2. We further discuss the limitations of both the evaluation and the tagger itself.

With a total of 74 ‘simple’ and over 70 additional ‘extended’ features covering fine-grained part-of-speech classes, tenses and aspects, frequent lexico-grammatical constructions, and numerous semantic categories, the MFTE’s feature portfolio meets the first criterion of identifying a broad range of lexico-grammatical and semantic features of English (see Table 1 in Section 2.2). Aside from the SO and LIKE features (see Section 2.4) they can readily be meaningfully interpreted (criterion 1a). Of the features in the simple tagset, all but twelve features reached “a satisfactorily high degree of accuracy” (see Table 1), with both precision and recall rates of > 90 per cent (criterion 1b). That said, it is worth remembering that some of the per-feature precision and recall rates presented in this study are based on relatively very few data points (see section 4.3).

Some of the most frequently mistagged features are among the problematic features for which some users of even the more recent versions of the *Biber tagger* have reported performing (semi-)automatic fix-tagging (e.g., features involving various uses of *that* and non-finite *-ed* and *-ing* clauses, see Goulart 2022 or Wood 2023). Low accuracy metrics

for some of these features may be the price of meeting specification criterion 1c: feature identification “without the need for human intervention” (see Section 2.2). For some features, we consider some of relatively low accuracy rates to be necessary to meet criterion 1a: if linguistic features are to be “meaningful” (in the sense of functionally and linguistically interpretable), it is crucial to distinguish, for example, non-finite past participle forms (VBN) from finite verbs in the perfect aspect (PEAS), and imperatives (VIMP) from infinitives (VB). Many existing taggers do not do this and therefore achieve higher accuracy rates overall, but the features are less readily interpretable. In sum, the results of the evaluation suggest that, at least for the simple tagset, most feature counts can be entered into multivariable analyses “without the need for human intervention” (1c).

While the MFTE struggled with e-language most (especially text messages and social media posts), the results of our evaluation nevertheless confirm that the MFTE performs sufficiently well across “a broad range of English registers” (thus meeting criterion 1d) with texts in “standard American or British orthography” (thus meeting criterion 1e). We recommend that researchers carry out additional evaluation tests if they intend to use the MFTE for other registers and/or varieties of English. Particular care should be taken when applying the extended tagset of the MFTE, as the accuracy of these features has not been subject to such a systematic evaluation.

The outputs of the MFTE also conforms to the tagger specification. Not only does the MFTE produce a table of raw counts (thus allowing researchers to apply their own normalisations) and two tables of normalised counts (criterion 2b), but it also saves the tagged texts for detailed examination of the texts themselves (criterion 2a). This is important to ensure full transparency of the tagging process and to verify the accuracy of the tagger’s output.

To meet the final requirement of the tagger specification, the MFTE source code and all the additional evaluation materials are available under a GPL-3.0 licence for use and scrutiny by the research community (criterion 3a). Criterion 3b is also satisfied with the publication of step-by-step instructions on how to install and run the MFTE on the landing page of its [GitHub repository](#), together with the present article describing the development and evaluation of the tagger.

6. CONCLUSION

In this article, we have presented the *Multi-Feature Tagger of English* (MFTE), an open-source tool designed for multivariable analyses of English. Characterised by transparency, adaptability, and accessibility, the MFTE offers promising avenues for future research endeavours in line with the principles of Open Science. Unlike ‘standard’ POS taggers such as the *Stanford tagger* (Toutanova *et al.* 2003) or CLAWS (Leech *et al.* 1994; Rayson and Garside 1998), its output does not require any additional processing (i.e., it outputs tables of counts with different normalisation options) and aims to tag only linguistically meaningful, functionally relevant features. As a free tool, the MFTE can contribute to making multivariable analysis of English more accessible to researchers and students from institutions with fewer resources.

While MAT and the initial version of the MFTE are in *Perl*, the new MFTE runs in *Python*, a programming language increasingly familiar to linguists. We hope that both the use of an accessible, object-based language and of an open-source licence will encourage colleagues not only to make use of the tagger, but also to contribute improvements. Our hope is that, as the tool gains traction within the research community, collaborative efforts will lead to further enhancements, expanding the tagset and refining the tagger’s performance across varieties and text types. Compared to the *MFTE Perl*, this new *Python* version benefits from being written in an actively developed language with a large user base, thus facilitating updates and improvements from developers (e.g., the integration of additional features using native parser libraries in *Python*).

In addition, the transparency and accessibility of the MFTE may also inspire linguists to develop similar taggers for languages other than English. Indeed, although MD analysis is, in theory, applicable to any language, in practice, most MD studies to date have examined varieties of English, which we believe is in part attributable to the lack of open-source taggers of lexico-grammatical features for languages other than English.

In discussing the results of the extensive evaluation of the MFTE’s simple tagset, we have also acknowledged the limitations of the tagger. Its accuracy will undoubtedly vary with respect to registers, topic domains, and varieties of English not included in the evaluation corpus. The extended tagset has not yet been systematically evaluated, and its reliance on dictionary lists for the semantic features is clearly a limitation. Refining and

updating these lists will be essential for the continued accuracy of the tagger. Future studies could explore the generation of tailored test data using Large Language Models (LLMs) as a means of evaluating the precision and recall rates of infrequent linguistic features. Finally, although the results of the evaluation are considerably more detailed than those of most linguistic taggers, they should nonetheless be interpreted with an awareness of the inherent challenges of making objective, categorical judgments when interpreting complex and often ambiguous linguistic phenomena. It would be misleading to suggest that these judgments are theory-free. As Gray (2019: 45) points out in an article focusing on the tagging and counting of linguistic features for MD analysis, “conflicting POS categorisation reflects a different grammatical interpretation or theory of the nature of this word.”

With these considerations in mind, we also hope that the MFTE will not only make a significant contribution to multivariable corpus linguistics research, but also stimulate ongoing methodological discussions on the transparency, validity, and reliability of the tools and methods used in corpus linguistics research. Ultimately, we hope that, in the near future, making research materials, data, and code available alongside linguistics publications will no longer be the exception (Wieling *et al.* 2018; Bochynska *et al.* 2023), but the norm.

REFERENCES

- Baker, Monya. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533/7604: 452–454.
- Barlett, Tom and Gerard O’Grady eds. 2017. *The Routledge Handbook of Systemic Functional Linguistics*. London: Routledge.
- Berber Sardinha, Tony, Marcia Veirano Pinto, Cristina Mayer, Maria Carolina Zuppari and Carlos Henrique Kauffmann. 2019. Adding registers to a previous multi-dimensional analysis. In Tony Berber Sardinha and Marcia Veirano Pinto eds. *Multidimensional Analysis: Research Methods and Current Issues*. New York: Bloomsbury, 165–188.
- Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton and Peter Pulsifer. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56/1: 1–18.
- Biber, Douglas. 1984. *A Model of Textual Relations within the Written and Spoken Modes*. California: University of Southern California dissertation.
- Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, Douglas. A typology of texts. *Linguistics* 27: 3–43.

- Biber, Douglas. 2006. *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.
- Biber, Douglas. 2019. Multidimensional Analysis: A historical synopsis. In Tony Berber Sardinha and Marcia Veirano Pinto eds. *Multi-Dimensional Analysis: Research Methods and Current Issues*. London: Bloomsbury Academic, 11–26.
- Biber, Douglas and Jesse Egbert. 2018. *Register Variation Online*. Cambridge: Cambridge University Press.
- Biber, Douglas and Bethany Gray. 2013. Discourse characteristics of writing and speaking task types on the TOEFL IBT test: A lexico-grammatical analysis. ETS Research Report Series 2013/1. <https://doi.org/10.1002/j.2333-8504.2013.tb02311.x>.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *The Longman Grammar of Spoken and Written English*. Harlow: Longman.
- Bochynska, Agata, Liam Keeble, Caitlin Halfacre, Joseph V. Casillas, Irys-Amélie Champagne, Kaidi Chen, Melanie Röthlisberger, Erin M. Buchanan and Timo B. Roettger. 2023. Reproducible research practices and transparency across linguistics. *Glossa Psycholinguistics* 2/1. <https://doi.org/10.5070/G6011239>.
- Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.
- Brezina, Vaclav, Abi Hawtin and Tony McEnery. 2021. The written British National Corpus 2014 – design and comparability. *Text & Talk* 41/5–6: 595–615.
- Davies, Mark. 1990. *Corpus of Contemporary American English* (COCA). <https://www.english-corpora.org/coca/>.
- Dixon, Daniel Hobson. 2022. *The Language in Digital Games: Register Variation in Virtual and Real-World Contexts*. Flagstaff: Northern Arizona University dissertation.
- Egbert, Jesse, Tove Larsson and Douglas Biber. 2020. *Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User*. Cambridge: Cambridge University Press.
- Egbert, Jesse and Shelley Staples. 2019. Doing multi-dimensional analysis in SPSS, SAS, and R. In Tony Berber Sardinha and Marcia Veirano Pinto eds. *Multi-Dimensional Analysis: Research Methods and Current Issues*. London: Bloomsbury Academic, 125–144.
- Gewin, Virginia. 2016. Data sharing: An open mind on open data. *Nature* 529/7584: 117–119.
- Goulart, Larissa. 2022. *Communicative Text Types in University Writing*. Flagstaff: Northern Arizona University dissertation.
- Goulart, Larissa and Margaret Wood. 2021. Methodological synthesis of research using multi-dimensional analysis. *Journal of Research Design and Statistics in Linguistics and Communication Science* 6/2: 107–137.
- Gray, Bethany. 2015. *Linguistic Variation in Research Articles: When Discipline Tells only Part of the Story*. Amsterdam: John Benjamins.
- Gray, Bethany. 2019. Tagging and counting linguistic features for multi-dimensional analysis. In Tony Berber Sardinha and Marcia Veirano Pinto eds. *Multi-Dimensional Analysis: Research Methods and Current Issues*. London: Bloomsbury Academic, 43–66.
- Gray, Bethany and Douglas Biber. 2013. Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics* 18/1: 109–135.
- Hidalgo, Bertha and Melody Goodman. 2013. Multivariate or multivariable regression? *American Journal of Public Health* 103/1: 39–40.

- In'nami, Yo, Atsushi Mizumoto, Luke Plonsky and Rie Koizumi. 2022. Promoting computationally reproducible research in applied linguistics: Recommended practices and considerations. *Research Methods in Applied Linguistics* 1/3. 100030. <https://doi.org/10.1016/j.rmal.2022.100030>.
- John, Leslie K., George Loewenstein and Drazen Prelec. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* 23/5: 524–532.
- Le Foll, Elen. 2021. *Introducing the Multi-Feature Tagger of English (MFTE)*. Perl. Osnabrück University. <https://github.com/elenlefol/MultiFeatureTaggerEnglish>.
- Le Foll, Elen. 2024. *Textbook English: A Multi-Dimensional Approach*. Studies in Corpus Linguistics 116. Amsterdam: John Benjamins.
- Le Foll, Elen and Muhammad Shakir. 2023. *Introducing a New Open-Source Corpus-Linguistic Tool: The Multi-Feature Tagger of English (MFTE)*. Paper presented at the 44th International Computer Archive of Modern and Medieval English Conference. NWU Vanderbijlpark: South Africa.
- Leech, Geoffrey, Roger Garside and Michael Bryant. 1994. CLAWS4: The tagging of the British National Corpus. In *Proceedings of the 15th conference on Computational Linguistics*. Kyoto: Association for Computational Linguistics, 622–628.
- Lindquist, Hans. 2009. *Corpus Linguistics and the Description of English (Edinburgh Textbooks on the English Language – Advanced)*. Edinburgh: Edinburgh University Press.
- 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, 171–189.
- McEnery, Tony, Richard Xiao and Yukio Tono. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. London: Taylor & Francis.
- McManus, Kevin. 2024. Replication and open science in applied linguistics research. In Luke Plonsky ed. *Open Science in Applied Linguistics*. Applied Linguistic Press, 148–165.
- Nini, Andrea. 2014. *Multidimensional Analysis Tagger (MAT)*. <http://sites.google.com/site/multidimensionaltagger>.
- Nini, Andrea. 2019. The multi-dimensional analysis tagger. In Tony Berber Sardinha and Marcia Veirano Pinto eds. *Multi-Dimensional Analysis: Research Methods and Current Issues*. New York: Bloomsbury, 67–96.
- Pashler, Harold and Eric-Jan Wagenmakers. 2012. Introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science* 7/6: 528–530.
- Picoral, Adriana, Shelley Staples and Randi Reppen. 2021. Automated annotation of learner English: An evaluation of software tools. *International Journal of Learner Corpus Research* 7/1: 17–52.
- Porte, Graeme and Kevin McManus. 2018. *Doing Replication Research in Applied Linguistics*. Milton Park: Routledge.
- Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv*. <https://doi.org/10.48550/arXiv.2003.07082>.
- Rayson, Paul and Roger Garside. 1998. The CLAWS web tagger. *ICAME Journal* 22: 121–123.
- Sinclair, John McH., Gwyneth Fox, Stephen Bullon, Ramesh Krishnamurthy, Elisabeth Manning and John Todd eds. 1990. *Collins Cobuild English grammar: Helping learners with real English*. Glasgow: Harper Collins.

- Sönning, Lukas and Valentin Werner. 2021. The replication crisis, scientific revolutions, and linguistics. *Linguistics* 59/5: 1179–1206.
- Toutanova, Kristina, Dan Klein, Christopher D. Manning and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In Marti Hearst and Mari Ostendorf eds. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Edmonton: Association for Computational Linguistics, 173–180.
- Wallis, Sean. 2020. *Statistics in Corpus Linguistics Research: A New Approach*. London: Routledge.
- Wieling, Martijn, Josine Rawee and Gertjan van Noord. 2018. Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics* 44/4: 641–649.
- Wood, Margaret. 2023. *Communicative Function and Linguistic Variation in State Statutory Law*. Flagstaff: Northern Arizona University dissertation.

Corresponding author

Elen Le Foll
 University of Cologne
 Department of Romance Studies
 Universitätsstraße 22
 50937 Cologne
 Germany
 E-mail: elefoll@uni-koeln.de

received: December 2023
 accepted: September 2024

Same, same, but *erm sort of* different? Comparing three kinds of fluencemes across Australian, British, Canadian, and New Zealand English

Karola Schmidt^a – Sandra Götz^b – Katja Jäschke^c – Stefan Th. Gries^{ad}

Justus Liebig University Giessen^a / Germany

Philipps University Marburg^b / Germany

University of Siegen^c / Germany

University of California, Santa Barbara^d / United States

Abstract – Although L1-English fluency has been extensively studied from many angles, few contrastive studies examine whether fluency develops similarly or differently across L1-varieties while taking sociolinguistic variation into consideration. This paper aims to close this research gap and examines the use of three core strategies of fluency (or fluencemes), i.e. discourse markers, filled pauses and unfilled pauses, across Australian, British, Canadian, and New Zealand English. These fluencemes were extracted and manually disambiguated from the private conversation sections of the respective components of the *International Corpus of English* (ICE-AUS, ICE-GB, ICE-CAN, and ICE-NZ). The data were normalised per speaker and linked with the sociobiographic metadata of the speakers. Analysis using random forests revealed a consistent fluenceme distribution across the four varieties, with unfilled pauses being the most common, followed by discourse markers, and then filled pauses. This pattern suggests a ‘common fluenceme core’ among L1-English varieties. The influence of sociolinguistic variables —gender, age, education, and occupation— was modest across varieties and exhibited diverse trends. Male speakers tend to use filled pauses more frequently but fewer unfilled pauses compared to female speakers. Increasing age did not significantly affect the frequency of these strategies; however, older speakers tend to use discourse markers less frequently. Both education and occupation showed a slight positive correlation with overall fluency.

Keywords – fluency; filled pauses; unfilled pauses; discourse markers; spoken English; inner-circle varieties of English



1. INTRODUCTION¹

Fluency in English has been widely researched in the past few decades, with previous studies having mainly taken psycholinguistic, cognitive, or sociolinguistic perspectives (e.g. Goldman-Eisler 1961; Albert 1980; Tottie 2015; Crible 2018; Beier *et al.* 2023). It is common knowledge that L1-speakers “violate the idealised rules of spontaneous speech while the discourse itself is still (most generally) perceived as fluent” (Dumont 2018: 63; see also Corley *et al.* 2007). Studies indicate that speakers usually bridge the gap between online processing demands and speaking by using different kinds of fluencemes, that is, “abstract and idealized feature[s] of speech that contribute [...] to the production and perception of fluency, whatever [...] [their] concrete realization might be” (Götz 2013: 8). Fluencemes can thus be used as strategies to overcome necessary planning phases in speech and can be realised, for example, as filled pauses (such as *er*, *erm*), unfilled pauses (i.e. pauses that are not filled with a non-verbal sound), discourse markers (such as *you know* and *like*, etc.), “smallwords” (Hasselgren 2002) (such as *sort of/sorta*, *kind of/kinda*), repeats (e.g. *it’s it’s it’s difficult*), incomplete utterances, etc. The necessity to use such planning strategies has been widely recognised in earlier research on fluency, for example by Beeching (2016: 100), who argues that “[p]ragmatic markers are perfectly adapted to the linear online editing which is required in spontaneous speech,” or by O’Connell and Kowal (see 2005: 572) in their work on filled pauses. Despite this general consensus about the frequent presence of fluencemes in L1-English speech, research has not yet revealed whether their frequencies and distributions are similar or different across L1-varieties of English. Accordingly, contrastive studies investigating such potential differences have only been rarely conducted, and, if so, mainly compared only two L1-varieties and/or only one fluenceme (e.g. Miller 2009 on discourse markers in Australian vs. New Zealand English, or Tottie 2011, 2015 on the use of filled pauses as ‘planners’ in British vs. American English). Moreover, regarding potential sociolinguistic variation in the use of fluencemes, previous research has predominantly concentrated on a single variety of English (e.g. Weiss *et al.* 2006; Laserna *et al.* 2014; Fruehwald 2016;

¹ We would like to gratefully acknowledge that this research project has been generously funded by the German Research Foundation (DFG, Reference Numbers GO 1760/4-1 and WO 2224/1-1) as part of a larger project on “Fluency in ENL, ESL and EFL: A contrastive corpus-based study of English as a first, second, and foreign language.” We would also like to thank Christoph Wolk for programming the data coding app and our student assistants Lara Möller, Hannah Vehrs, and Daniel Walker for their invaluable help with the data coding and their patience with long discussions over fluencemes. All remaining errors and infelicities are, however, our responsibility alone.

Scheuringer *et al.* 2017; Sokołowski *et al.* 2020; the studies in Leuckert and Rüdiger 2021). Whether or not there is sociolinguistic variation in fluenceme use across L1-varieties is still largely unknown. Furthermore, while there is a substantial body of research on the use of individual fluencemes, it tends to focus heavily on either discourse markers or filled pauses (see Section 2). Studies exploring fluencemes in combination, however, remain notably scarce. To the best of our knowledge, no contrastive research has yet examined the combination of different fluencemes across English varieties while also addressing sociolinguistic variation.

In this paper, we aim to advance the cumulative understanding of fluency by addressing these research gaps. In doing so, we examine how L1-English speakers from four L1-varieties of English use fluencemes to navigate planning phases, considering a range of social variables. Accordingly, this paper seeks to address the following research questions:

- 1) Are there differences in fluenceme use between different L1-varieties of English?
- 2) Do sociolinguistic variables play a role in predicting the choice of particular fluencemes across varieties?

The remainder of this paper is thus structured as follows. After giving an overview of previous research on fluency and fluencemes in L1-English (Section 2), we present the database used and methods applied (Section 3). In Section 4, we present the findings of our analysis and then round up this paper with a discussion, a conclusion, and an outlook to future research in Section 5.

2. FLUENCY IN L1-VARIETIES OF ENGLISH

Although fluency boasts a relatively long research tradition in linguistics, there are considerable differences in how the concept is defined and operationalised. There seems to be a consensus that fluency is linked to swift and effortless speech production (see Chambers 1997: 535). On a conceptual level, fluency is, however, more frequently linked to learner language than to L1-speech and, therefore, is often paired with the concept of accuracy (Chambers 1997: 536; see also Brand and Götz 2011). Viewed as a concept intertwined with accuracy, the approach of applying fluency to L1-speech may seem absurd. Fluency understood as a marker for language proficiency is not applicable to L1-speakers, which is why Riggenbach observes that “in common usage rarely does one hear

a native speaker being called fluent in comparison to other native speakers” (1991: 424). If someone’s first language is, for example, British English, speakers are simply considered as being fluent in English and the same perception is true for speakers of other L1-varieties of English.

However, if one’s view of fluency is broadened and detached from the concept of accuracy, it can also be used to describe and analyse L1-speakers’ utterances. Producing speech always necessitates bridging the gap between thought and perceivable language. Given the rapid pace of oral speech production, this poses a challenge even for highly proficient speakers. Segalowitz labels this as “utterance fluency,” that is, “the features of utterances that reflect the speaker’s cognitive fluency” (2010: 165). Utterance fluency can be operationalised as the total or relative amount of fluencemes a speaker uses. Accordingly, in our understanding, we do not consider such strategies negatively connoted markers of disfluency (e.g. Maclay and Osgood 1959), but rather as fluency-enhancing strategies.

2.1. Utterance fluency in L1-varieties of English

There are some corpus-based studies on fluency in L1-varieties of English such as Götz (2013), Osborne (2013), Crible (2018), and Dumont (2018). Although these studies’ set-ups slightly differ from ours with, for example, Crible (2018) and Dumont (2018) including conjunctions and Osborne’s (2013) focus lying on temporal fluency and syntactic patterns, their findings are still relevant for our study. Crible, for example, reports that “fluencemes are omnipresent in speech production, covering about one fifth of the sound signal, and such momentary interruptions of the smooth unfolding of speech mostly attend to the upcoming rather than the previous material” (2018: 133). The sheer rate of fluencemes found in Crible’s data already indicates that they “are the normal accompaniment of [...] speech” (Dumont 2018: 63). Fluencemes require further research into how they are distributed and what factors influence them. For this perspective, it is useful to take a closer look at the different categories of fluencemes that will be analysed in Section 4. When describing such fluency-enhancing devices, there are different labels to choose from. In this paper, we will adopt the terms ‘filled pauses’, ‘unfilled pauses’ and ‘discourse markers’ to refer to the three fluencemes we investigate, which are the three most frequently occurring strategies in speech, and thus we consider them to be ‘core fluencemes’.

2.2. Previous research on *fluencemes* in L1-English

Discourse markers have taken centre stage in previous research on L1-English (e.g. Brinton 1996; Aijmer 2002; Beeching 2016; Crible 2018, to name but a few). There has been extensive work on the social and pragmatic functions of discourse markers. Generally, there is consensus that discourse markers have no effect on the truth condition of an utterance, but Fung and Carter (2007: 414) point out that omitting them also means omitting clues for the listener as to how the truth condition is to be interpreted. Many different functions have been attested for discourse markers, for example, acting as planners, floorholders, and repair markers (Aijmer 2002: 51), as well as facilitating ease of listening for the reader (Aijmer 2002: 3), or as discourse initiators, boundary markers, floor managers, indicators of information status or to make implicit information explicit, functioning as responses to previous discourse or as hedging devices (see Brinton 1996: 37–38), as well as marking boundaries of talk and ending topics (see Fung and Carter 2007: 412). For the purposes of the present study, however, their varying functions beyond planning are of somewhat secondary importance, although discourse markers are not restricted to one discourse function alone and can potentially fulfil several of their attested functions at the same time. Beeching (2016: 99), for example, argues that the discourse marker *you know* is used for hesitation while being directed at the listener and asking the interlocutor to collaborate in bridging the planning phase.

Specific functions aside, discourse markers have been reported to “vary in adult native varieties depending on social context” (Fuller 2003: 192). In New Zealand English, for example, social class has been identified as a determining factor for the use of discourse markers such as *eh* and *you know*, which have been identified as features of the vernacular (see Stubbe and Holmes 1995: 74). However, the same study also concludes that New Zealand English shows a comparably small class effect in contrast with other varieties (1995: 72–73).

In addition to this kind of variety-internal variation, there are also differences in discourse marker use across different varieties. Speakers of New Zealand English use the discourse marker *like* significantly less frequently than their Australian counterparts, at least in private conversations. In Australian data, the use of *like* varies by the speaker’s age: while teenagers use *like* frequently, a decrease in frequency can be observed when speakers approach adulthood. This effect, however, cannot be found in New Zealand English where *like* is not age-restricted (see Miller 2009: 327–328), indicating an age-by-

variety interaction. In a similar vein, Reichelt (2021) shows that discourse markers can also vary across social factors. She finds that the choice between the discourse markers *kind of* and *sort of* is sensitive to age, amongst other factors. Other sociolinguistic variables have also been attested to influence discourse marker usage (e.g. Tagliamonte 2005 on gender; Fuller 2003 on speech context). In addition, Buysse and Blanchard (2022) demonstrate that L1-English speakers have an overall positive attitude towards discourse markers, particularly associating them with friendliness and confidence. In fact, even though discourse markers tend to be marked as signals of a low education level and lack of politeness, the L1-speakers almost invariably rated speech with a frequency of discourse markers much higher than their L2 counterparts (see Buysse and Blanchard 2022: 239–240), suggesting that discourse markers are an accepted asset to L1-speech. Thus, there is a solid foundation of research showing that discourse markers are inextricably linked to their social context. Although we only focus on the planning function of discourse markers, we also expect to see connections to sociolinguistic variables when we take a comparative approach to fluency across varieties of English.

The second most frequently researched fluenceme is filled pauses, such as *er/uh* or *erm/urm*. Filled pauses' prolific research history seems natural, given that they are reported to make up 6.38 per cent of speech (see Clark and Fox Tree 2002: 81). Filled pauses are particularly important in fluency research, because they are closely connected to a speaker's utterance fluency (e.g. Clark and Fox Tree 2002; O'Connell and Kowal 2005, among many others). Previous research has attested that filled pauses can be considered as words or 'true' linguistic items that follow clear usage patterns (e.g. Kjellmer 2003; Tottie 2011, 2015). Kirjavainen *et al.* (2022) support the classification of filled pauses as grammatical items. Their findings suggest that, at least in British English, the placement of filled pauses is relatively constrained. They also find that listeners typically do not notice the presence of filled pauses, similar to grammatical structures. Consequently, they tentatively suggest that speakers could be processing filled pauses in the same manner as grammatical structures (Kirjavainen *et al.* 2022).

A large body of research into filled pauses was contributed by Gunnel Tottie. Her work on filled pauses ('planners' in her terminology; Tottie 2011: 193) includes contrastive findings on two L1-varieties of English, namely British and American English. She finds that filled pauses display a higher frequency in British than in American English, and she also attests differences in other usage patterns, for example,

in the co-occurrence with unfilled pauses and the distribution of nasalised and non-nasalised variants of filled pauses (Tottie 2015: 42, 47, 49). Tottie (2011) also identifies several sociolinguistic factors as significant predictors of filled pause usage. In her data, men produce more filled pauses than women (2011: 182). Additionally, filled pauses are most frequent among people who are 60 years or older, although the data do not clearly indicate an overall age-grading effect, given that there is no discernible trend in other age groups (2011: 186–187). Lastly, her data show social class to play a significant role. Speakers from what she identifies as the highest socio-economic class produce significantly more filled pauses than other speakers (2011: 188).

Although there is some debate about the terminology of filled pauses and their exact functions (see, e.g., Revis and Bernaisch 2020: 137–138), with some perception studies even suggesting that filled pauses are used more frequently when speakers are lying (e.g. Loy *et al.* 2017), they occur relatively frequently in their function as fluencemes (MacLay and Osgood 1959: 34; Götz 2013). Their other functions notwithstanding, it stands to reason that filled pauses in particular count towards phenomena that are used systematically to overcome planning phases. O’Connell and Kowal arrive at the same conclusion and assert that *uh* and *uhm* “protect and sustain genuine fluency” (2005: 527).

The third core fluenceme we are concerned with in the context of the present study is unfilled pauses. Their occurrence in speech is considered to be a natural (Biber *et al.* 1999: 1054) and non-random phenomenon (see Svartvik 1990: 73). While some unfilled pauses are placed intentionally by the speaker to fulfil clear communicative functions (Drommel 1980) and are perceived as being natural by the listener (Chafe 1980), others occur during breathing or gesturing, etc. However, the majority of unfilled pauses occur for speech-planning reasons (Drommel 1980). They have been shown to occur most frequently at major grammatical transition points, for example, between syntactic units or at points where utterance launchers (such as *oh*, *well*, *okay*) are likely to occur (see Biber *et al.* 1999: 1054). In a similar vein, Lounsbury (1954) distinguishes between what he calls ‘juncture pauses’ at major syntactic boundaries, and ‘hesitation pauses’ within syntactic units, where listeners are more tolerant to the former and perceive the latter as being longer (Butcher 1980). Pausing within constituents and towards clause boundaries has been shown to be quite common and natural in speech (e.g. Pawley and Syder 1983: 200; Götz 2013; Revis and Bernaisch 2020: 139). MacLay and Osgood (1959: 31) have found that L1-English speakers tend to use unfilled pauses before beginning a phrase and

then before uttering the content word in that phrase. This pattern of placing pauses at mid- and end-clause position within an utterance is one feature of what is labelled “breakdown fluency” by Foster and Tavakoli (2009: 875). Unfilled pauses occur with such high frequencies in their function as fluencemes that they can be seen as good indicators of a person’s overall productive/utterance fluency (see Goldman-Eisler 1961; Segalowitz 2010). That being said, not much is known about unfilled pauses and their usage in correlation with sociolinguistic variables.

As our literature review reveals, most research conducted so far presents itself in the form of (case) studies of small sets of fluency-enhancing devices of individual fluencemes on individual discourse markers. Studies covering different varieties include BurrIDGE and Florey (2002), Miller (2009), BurrIDGE (2014), Vine (2016), and Burke and BurrIDGE (2023) on discourse markers in Australian and New Zealand English respectively, as well as Grant’s (2010) comparison of *I don’t know* in British and New Zealand English. Canadian English appears to be rather under-researched in terms of fluency, although there has been research on individual discourse markers, for instance, Tagliamonte’s (2005) study on the use of *so*, *like* and *just*. These comparative studies indicate that there is some degree of variation between fluencemes across different varieties of English. It has also been shown that sociolinguistic variables can influence fluenceme use (e.g. Bortfeld *et al.* 2001 for American English). We therefore assume to find differences in fluenceme use to be caused by (an interplay of) different varieties of English and sociolinguistic variables, such as gender, age or occupation.

2.3. *Towards a contrastive approach to fluency across L1 varieties of English*

The paper at hand is concerned with British English (BrE), Australian English (AusE), New Zealand English (NZE), and Canadian English (CanE).² We chose to work with these four L1-varieties as we wanted to compare differences between varieties spoken on different continents and in countries in which they constitute one of the dominant languages or even the dominant or official language (Kachru 1985: 12). Because these

² While it would have been desirable to include American English in our analysis, we opted for Canadian English instead, because it is the geographically closest neighbour to American English for which we had the same type of data as for the other varieties. Since fluency has been shown to be very register- and text-type sensitive, we wanted to make sure to only compare data that were as similar as possible.

varieties have been extensively researched with respect to regional or social variation within each variety itself, we will not describe them further at this stage.

All these four Englishes under scrutiny but BrE are severely under-researched in terms of fluency in general and in terms of comparative fluenceme use in particular. Against this backdrop, this paper takes a first step toward addressing this research gap by examining potential differences in fluenceme use across four varieties of English. We also incorporate sociolinguistic factors into our analysis to explore their role in determining the choice of specific strategies. We also test for interactions between these variables as it might be the case, for example, that gender affects fluency performance, but only for speakers of a certain variety.

3. DATABASE, CODING PROCEDURE, AND APPLIED METHODOLOGY

3.1. Database and coding procedure

We selected four components of the *International Corpus of English* (ICE), namely the BrE (ICE-GB), AusE (ICE-AUS), NZE (ICE-NZ), and CanE (ICE-CAN) components. The ICE corpora are highly suitable for our purposes, as all ICE components were compiled using the same design and applying common schemes for annotation, which ensures comparability across different sub-corpora (Greenbaum and Nelson 1996).³ Each component contains 300 spoken text files of approximately 2,000 words per file from different types of settings, such as dialogues and monologues, private, public, scripted, and unscripted speech. We chose 100 texts containing unscripted private dialogues from either face-to-face or telephone conversations, because we consider these data to be the most natural data type in the spoken section of the corpus.⁴ Our database thus comprises *ca.* 200,000 words per variety and *ca.* 800,000 words in total. Managing and coding such a large amount of data was possible by taking several steps. We initially identified relevant fluencemes by carefully reviewing the pertinent literature and manually adding further items to the categories under investigation (i.e., unfilled pauses, filled pauses, and discourse markers). For each subcorpus, we manually added variety-specific fluencemes.

³ Of course, there are some remaining issues with regard to the comparability of the different components of the ICE components, but we consider them to strike the best balance in terms of comparability and diversity of registers/genres represented in the varieties' components.

⁴ Obviously, the analysis proposed here is based on the assumption that the corpus annotation is sufficiently precise and attempts to benefit from the fact that, while certainly less precise than acoustic measurements, it at least involves many more speakers than such analyses can reasonably be expected to include.

For example, in ICE-NZ we found that the items *yeah no* and *yeah nah* are used as fluency enhancing devices (see also BurrIDGE and Florey 2002; Manhire 2021). We added such kinds of frequently occurring discourse markers so as not to miss variety-specific fluencemes which might be used as alternative strategies to other discourse markers.

The identified fluencemes were uploaded to a coding application developed by Wolk *et al.* (2021) in *R* (R Core Team 2024) utilising the web application framework *shiny*.⁵ This tool is able to identify the uploaded items via string search in the text files of the corpora under investigation. The tool's interface displays the transcribed conversations, which are only available as unformatted text files, as dialogues and highlights the identified items. As the tool identifies strings and not syntactic categories, it is necessary to manually disambiguate the detected items. This is illustrated in the examples in (1). In (1a), *like* is a verb, not a fluenceme, but in (1b) *like* can be considered to be a fluency-enhancing device.

(1a) I *like* his new car.

(1b) This is *like* really new information.

Given the extreme variability of spoken language, the sometimes-erroneous transcriptions, and missing information on intonation, the process of manual disambiguation is extremely challenging, so we took several measures to ensure reliable annotation. A coding manual was issued and constantly revised and adapted during the coding process and we trained our coders using training files and coded every file in several steps. First, each file was initially coded by a student assistant. A second student assistant reviewed the entire file for coding accuracy, noting any discrepancies and discussing them with the original coder. Items on which agreement could not be reached were referred to a senior researcher for resolution. If the senior researcher was also uncertain about the item's status, it was brought before the entire team of four student assistants and three senior researchers. A majority vote was used to decide if consensus could not be reached. Truly ambiguous cases were deleted from the dataset. We deemed this four-step coding process superior to a double-blind coding approach that merely calculated annotator deviations, as it allowed us to carefully discuss the contexts of any potentially ambiguous fluencemes. This ensured that our dataset consisted solely of truly

⁵ More information about *shiny* is available at <https://shiny.rstudio.com>.

accurate hits. The coding guidelines we use will be summarised in the following with reference to exemplary problems that occurred during the annotation process.

3.1.1. Discourse markers

Inspired by constituency tests which can be found in most syntax textbooks (e.g. Börjars and Burridge 2010; van Gelderen 2010), we consider discourse markers as fluencemes if (i) they can be omitted without either changing the truth value of an utterance, (ii) if they can be omitted without rendering the utterance ungrammatical, or (iii) if they can be replaced by another fluenceme or another discourse marker. Although these simple tests worked for a large part of the dataset, numerous instances remained that were difficult to code. One example is potential fluencemes that occur in utterance-final position. Theoretically, an utterance-final *like* can more often than not be a fluenceme or also fulfil another syntactic function, as shown in example (2).

- (2) Speaker A: I started university this year. It is *like*...
 Speaker B: What do you study?

As Speaker A was interrupted by Speaker B, it is impossible to tell what exactly Speaker A wanted to say. For example, it could be *It is like school*, in which *like* does not act as a fluenceme, or *It is like really boring*, where it has clear discourse marker status. Thus, we chose to not include such pre-interruption items, as well as all other cases that were truly ambiguous. Altogether, we annotated and disambiguated 33 different types of discourse markers and smallwords that were used in a planning function, as illustrated in Table 1.

3.1.2. Filled and unfilled pauses

Filled pauses function as fluencemes in most cases, but still needed to be disambiguated. Speakers especially in NZE and AusE, for example, also frequently use these sounds as question tags (e.g. *You saw that, eh?*), either to react to an utterance or as a backchanneling response (e.g. Schweinberger 2018). Since such instances do not serve the fluency-enhancing function we are investigating, they were excluded from our dataset, as well as all other functions in which the filled pauses listed were not used as planning devices (e.g. backchannels, emphasisers, etc.). If it was impossible to determine their function, they

were deleted from the dataset. The set of filled pauses analysed in our study consists of 19 types, which are illustrated in Table 1.

Fluenceme	Subtypes
Discourse markers	<i>actually, alright, alright, anyhow, anyway, basically, do you know what I mean, huh, I don't know, I mean, in a way, kind of, kinda, know what I mean, let's see, like, nah, no, okay, or something, right, so, sort of, sorta, stuff like that, thing like that, things like that, well, yeah, yep, yes, you know, you see</i>
Filled pauses	<i>ah, ahh, ahm, ahn, eh, ehm, er, haan, hmm, mm, mmm, mhm, uhm, uhn, hm, ur, urm, uh, uhh</i>
Unfilled pauses	<i><, >, <., ></i>

Table 1: Overview of investigated types of fluencemes

Finally, unfilled pauses are automatically accepted as fluencemes by the corpus tool and are not manually disambiguated.

3.2. Clean-up and annotation of the dataset

Following the disambiguation, the resulting fluenceme data were further cleaned. Crucially, as we study L2-English speakers, all speakers who had indicated a first language other than English were removed from the dataset. As the central point of the analysis, we then conflated the fluenceme categories. In the coding tool's initial output, each fluenceme is categorised in the fine-grained manner detailed in Table 1 above and Table 2 below so that, for example, one can distinguish between the different realisations of discourse markers. For the present analysis, we adopted a broader three-way categorisation with the levels 'discourse marker' – 'filled pause' – 'unfilled pause'. Next, we computed individual fluenceme counts for each speaker with attested fluenceme usage in the data. To this end, we used the corpus tags to automatically identify in each file any non-corpus material (e.g. insertions from the transcribers), which was then removed in order to compute the number of words uttered by each speaker. Using these word counts per speaker per file, we computed the response variable, that is, three normalised frequencies of fluencemes (one per fluenceme type) for each speaker, by dividing the absolute frequency of each fluenceme type by the number of words per speaker. Due to the nature of this normalisation, we were unfortunately not able to include the linguistic context of the fluencemes (e.g. the frequency or complexity of the lexeme following a fluenceme, or whether or not it was primed) to our analysis, although we believe that adding such 'level-1' variables (see Gries 2024) at a later stage could add more explanatory power to our analysis.

Lastly, we added (speaker-specific) information from the corpora’s metadata to be included in the analysis. Firstly, we entered both the variable `CORPUS` and the text category variable `TEXTCAT` for each speaker, that is, whether their data come from a face-to-face conversation or a phone call. Next, we annotated the speakers’ `GENDER` (i.e. *male* or *female*) as specified in the metadata. The metadata also included information on the speakers’ `AGE`, but with different binning between the four corpora. To account for this, we kept all individual ages (i.e. non-binned ones) as is and transformed all age ranges to their median. All four sets of metadata also provided information on speaker education and speaker occupation, which we also needed to homogenise, because the information given was too diverse (i.e. 94 different levels for education and 242 levels for occupation). We therefore used the format inspired by the scaling of indices related to the Pisa index of economic, social and cultural status (OECD 1999). For the speakers’ educational level `EDUCATION_COARSE`, we applied the format of the International Standard Classification of Education ISCED-97 (UNESCO 2006) and transformed the speakers’ education to a Level from 0–6, where 0 corresponds to elementary school and 6 to postgraduate education, such as a PhD (OECD 1999). Each occupation, as stated by the speakers in the corpus, was entered into a variable `OCCUPATION_ISCO880` transformed to the value corresponding to the International Standard Classification of Occupations (ISCO-08 (COM), ranging from 16 (e.g. farmers) to 85 (e.g. medical professionals/jurists) (Ganzeboom 2010).

Lastly, we added each `SPEAKER` identifiable by a unique ID as a variable to the dataset. This led to us having 51,950 data points for which all of the above-listed predictors available distributed across 719 speakers, as summarised in Table 2.⁶

	AustrE	CanE	BrE	NZE	Sum
Discourse markers	4,854	6,368	3,745	101	15,068
Filled pauses	1,966	3,444	2,713	48	8,171
Unfilled pauses	11,012	10,771	6,884	44	28,711
Sum (# of speakers)	17,832 (197)	20,583 (248)	13,342 (265)	193 (9)	51,950 (719)

Table 2: Overview of the final data

⁶ We are not providing a more detailed descriptive statistics breakdown per fluenceme and/or variety because these cannot do justice to the multifactorial nature of the data: mean frequencies, etc. grouped by only one predictor *per definitionem* come with a huge loss of information because they are aggregated over all other predictors.

3.3. Statistical analysis

Two kinds of analyses are possible for this kind of data, which come with different implementations:

- a type 1 analysis, in which the response variable is the fluenceme type (i.e., ‘discourse marker’ vs. ‘filled pause’ vs. ‘unfilled pause’) and where one would try to predict from the other variables what disfluency will be chosen in every single utterance case/context;
- a type 2 analysis, in which the response variable is the fluenceme type frequency (i.e. a numeric value of 0) and where one would try to predict the frequency with which speaker groups as defined by the other variables use which fluenceme how often.

Type 1 analyses are more common in corpus studies of variationist phenomena and might seem like the natural choice here. However, most such studies focus on hypothesis testing and include level-1 predictors (Gries 2024) that vary with the linguistic choice being studied. In contrast, our exploratory study aligns better with a kind of type 2 analysis focusing on how speaker groups, defined by annotated variables, tend to favour certain fluencemes. Thus, we adopt type 2 analysis here, reserving type 1 for future research. Type 2 analyses with response variables that are frequencies can be analysed in many ways, including various kinds of regression models, but we chose the tree-based method of random forests, which can be applied to frequencies without link functions, does not make any assumptions about the data,⁷ can handle potentially collinear predictors slightly better than regressions, and is extremely good at detecting interactions, that is, “identifying ‘the interaction effect per se and the predictor variables interacting with each other’” (Gries 2021: 481). The random forests approach is based on fitting many different classification or regression trees on a data set (e.g. 500), but adding two layers of randomness:

- each of the trees in the forest is fitted on a different randomly sampled-with-replacement version of the data;

⁷ Even count-specific regression models have assumptions (Poisson regression requires equal expectation/mean and variance, while negative binomial regression needs a link parameter) that tree-based methods, including forests, do not have.

- in each tree, only a randomly-selected subset of the predictors is eligible to be chosen for a split (see Gries 2021: 463–464).

These two characteristics are desirable inasmuch as they help to reduce the risk of overfitting (because trees are fit on many different versions of the data and because the prediction accuracy is evaluated against the data that did not make it into a specific sample) and in how they decorrelate trees (e.g. as very powerful predictors are less able to dominate every single tree because they might not always be available for a split). In addition, random forests (i) can be less sensitive to distributions that pose problems for regression, (ii) are good for small n -large p problems (i.e. scenarios with many predictors but not many data points), and (iii) often outperform regressions when it comes to making accurate predictions (e.g. because they can be more powerful at detecting non-linearity).

The main disadvantage of random forests is the challenge they pose for interpretation: unlike in a regression, there are no simple coefficients for differences between means or slopes, and unlike in a tree, there is no simple visualization of potentially thousands of trees in a forest. In addition, detecting interactions (i.e. identifying the interaction effect of 2+ predictors) is not as straightforward as is widely believed (see Wright *et al.* 2016; Gries 2020, 2021: 481–483). To shed light on the inner workings of the fairly black-box random forest, we use

- **variable importance scores** (permutation-based); note, these indicate “the impact of each predictor variable individually as well as in multivariate interactions with other predictor variables” (Strobl *et al.* 2009: 337);
- **partial dependence scores**; these indicate what in a regression model would be the direction/sign of effects of predictor values/levels. Note that, to identify interactions between predictors, we computed partial dependence scores not just for predictors on their own, but also for selected interactions of interest.

Even though random forests are already very powerful and flexible by default (the trees they are based on do not make distributional assumptions), they sometimes can benefit from at least some minor degree of preparation. In our case, the normalised frequencies of each fluenceme per speaker were very right-skewed. While the values ranged from 0 to 0.25, more than 80 per cent of them were below 0.05, so we applied a Box-Cox

transformation to avoid a long underpopulated tail on the right;⁸ as a result, the new version of these normalised frequencies was very symmetric.

The random forest implementation we chose for this study is the one implemented in the *R* package `{ranger}` (Wright and Ziegler 2017), which also provides the variable importance scores; the partial dependence scores were computed with the *R* package `{pdp}` (Greenwell 2017) and we followed a logic similar to the one discussed in Gries (2021: 480–483). The formula submitted to `ranger::ranger` was the one shown below and we built `ntree=3,000` trees with the default settings of `ranger`’s hyperparameters `mtry=2`, a target node size of 5, and permutation-based local variable importance scores.

```
NORMFREQ_bnc ~
  FLUTYPE + # discourse marker vs. filled pause vs. unfilled pause
  CORPUS + # Austr vs. Canadian vs. British vs. New Zealand
           English
  TEXTCAT + # direct vs. distance conversations
  SPEAKER + # the unique ID of each speaker (≅ a random effect)
           # and several speaker-related variables:
  AGE + GENDER + EDUCATION_COARSE + OCCUPATION_ISC088
```

4. RESULTS

Given that the response variable was only a normalised frequency and, thus, disregarded any contextual level-1 variables (see Gries 2024), the out-of-bag prediction accuracy was surprisingly high with an R^2 of 0.492. Figure 1 shows the relationship between the observed and predicted fluenceme frequencies with the modelled transformed frequencies on the bottom and left axes and the back-transformed frequencies on the upper and right axes.

⁸ We used the function `car::bcnPower` (Fox and Weisberg 2019) to estimate the best lambda (-2.094296) and the best gamma (0.1) values for the response variable.

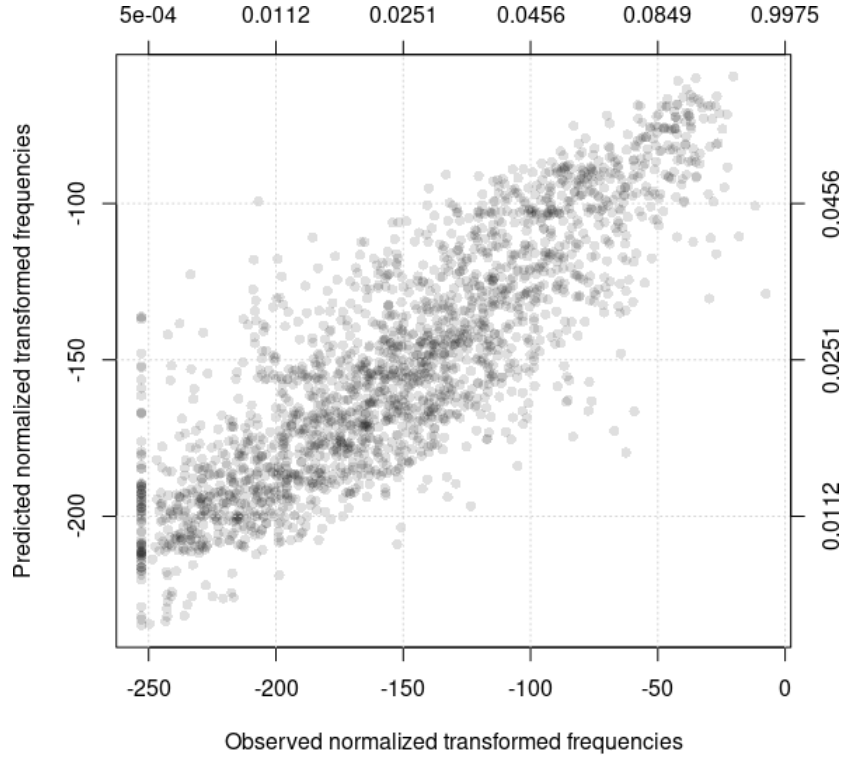


Figure 1: Predicted vs. observed proportions of fluenceme frequency

The variable importance scores returned the following order of predictors: FLUTYPE (2,519) >> SPEAKER (648) > CORPUS (469) > AGE (266) > OCCUPATION_ISCO88 (199) > EDUCATION_COARSE (78) \approx TEXTCAT (73) \approx GENDER (67). But the more interesting questions are of course how the most important predictors affect the prediction of the response—alone (as main effects) or in interaction patterns—and in which direction they do so. We therefore computed partial dependence scores (back-transformed to the original proportions for interpretability) for the predictors FLUTYPE, SPEAKER, CORPUS, and EDUCATION_COARSE, as well as for the ‘interactions’ FLUTYPE:CORPUS and AGE:GENDER(:FLUTYPE).

4.1. The effects of FLUTYPE, CORPUS and FLUTYPE:CORPUS

As might be expected from the importance scores, the effect of the predictor FLUTYPE is fairly strong: unfilled pauses are by far the most frequent (0.04), followed by a large margin by discourse markers (0.023), and filled pauses (0.014), as shown in Figure 2.⁹

⁹ While some researchers might object to the use of lines connecting the dots (given that, indeed, there are no data between, say, discourse markers and filled pauses), we still prefer using lines because of how they help guide readers’ visual processing of the relations between data points (see also Fox and Weisberg’s 2019 effects package).

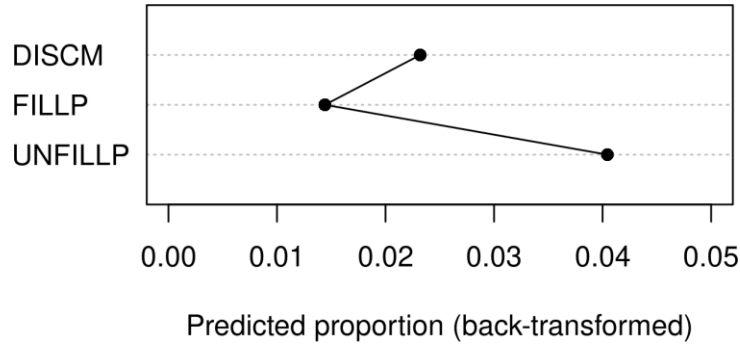


Figure 2: Predicted proportions of fluencemes by FLUTYPE

The effect of CORPUS is somewhat weaker and is largely driven by how NZE speakers behave differently from the rest: Australian, Canadian, and British speakers are quite similar (0.026, 0.026, and 0.024, respectively) whereas the NZ speakers have a notably lower frequency of 0.016 (see Figure 3). However, this effect of CORPUS must not be overinterpreted because (i) NZE was only represented by nine speakers in our data, and (ii) all of those were from one education level (the second highest).¹⁰

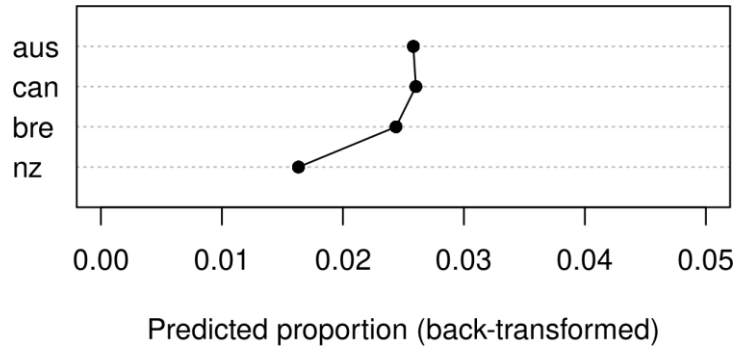


Figure 3: Predicted proportions of fluencemes per CORPUS

Given our interests, it was worthwhile to determine whether (part of) the reason for the high variable importance scores was an interaction of these two predictors. We therefore computed partial dependence scores for the interaction, both perspectives of which—in terms of which predictor is grouped by which other one—are shown in Figure 4.

¹⁰ The degree to which the NZE speakers are different from the speakers in the other varieties might lead some to suggest leaving them out, but we would consider that a mistake. We think it is better to provide the (limited) descriptive information they afford in the forest together with the context required to (not over)interpret the results than it is to simply pretend those data do not exist. For example, now that we have described the data to at least some extent, follow-up studies can use our description as a launchpad to motivate further data collection or to drill down into only the highest education levels of the other varieties.

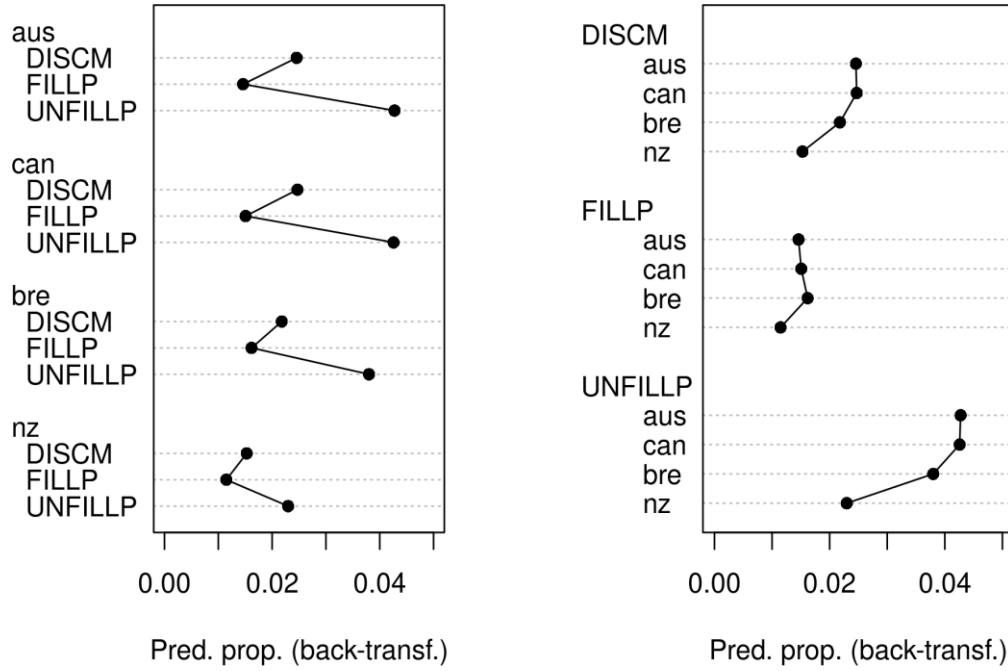


Figure 4: Predicted proportions of fluencemes for FLUTYPE:CORPUS

It seems that there is only a weak interaction. In all four varieties, unfilled pauses are much more frequent than discourse markers, which are more frequent (but less so) than filled pauses. However, NZE is peculiar because it features a much smaller number of unfilled pauses than any other variety. From a different perspective, we can see that

- the frequencies of discourse markers and unfilled pauses across varieties behave similarly to the frequencies of all fluenceme types across varieties;
- the frequencies of filled pauses vary least across varieties.

4.2. The effect of *SPEAKER*

There is a great degree of speaker heterogeneity, but given the high number of speakers and the lack of interpretability of individual speaker preferences, this can only be shown heuristically, as we do in Figure 5.

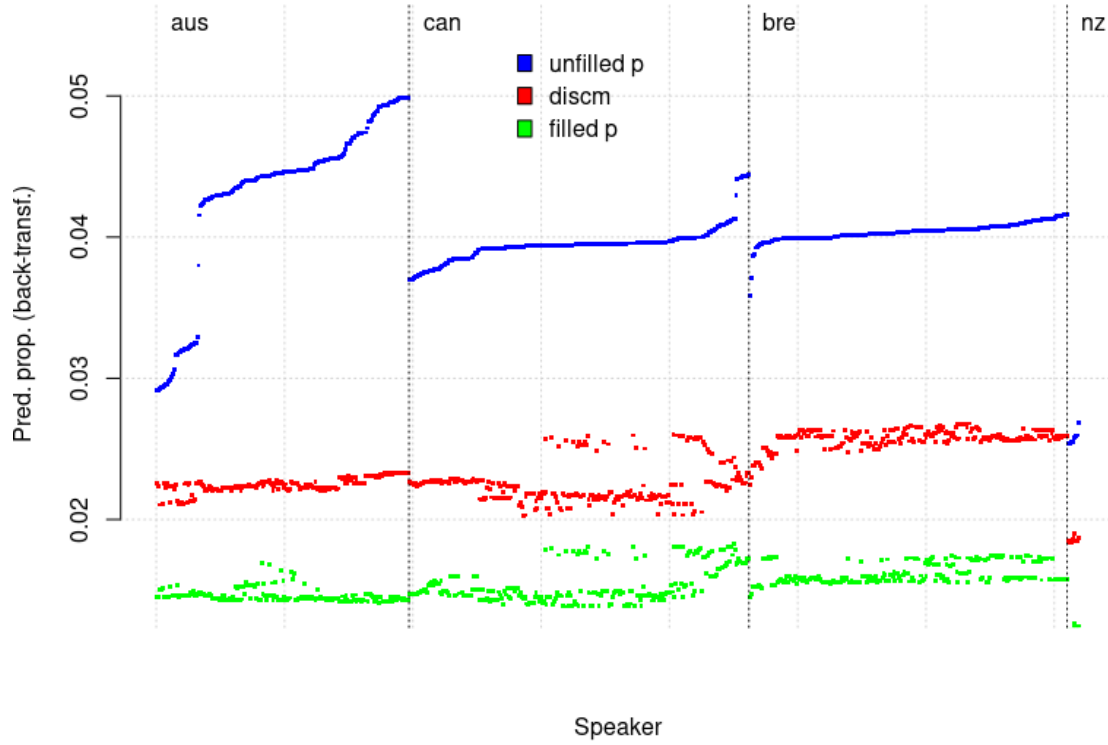


Figure 5: Predicted proportions of fluencemes per FLUTYPE:VARIETY/SPEAKER

The y-axis represents the predicted proportions of fluencemes with each fluenceme type in a different colour and the x-axis representing all 719 speakers grouped by variety. Note that the x-axis should not be interpreted sequentially, as it is not a variable like TIME; instead, one must interpret it vertically such that each position on the x-axis represents one speaker's three fluenceme frequency predictions. Most visually striking is of course the huge variability we observe for the unfilled pauses (in particular for the AusE speakers), but it is also worth noting that, for discourse markers and filled pauses, variation coefficients¹¹ indicate that the BrE speakers are most variable.

4.3. The effects of AGE, GENDER, AGE:GENDER, and AGE:GENDER:FLUTYPE

The two variables AGE and GENDER do not have particularly strong effects, but we were ultimately also interested in determining (i) whether the genders differed in their overall fluenceme differences, (ii) whether there was an age effect on overall fluenceme use, and (iii) whether any of these effects interacted. While this was not particularly likely, given the low variable importance scores of AGE and GENDER, we therefore computed partial

¹¹ For unfilled pauses (in blue), we obtained AusE (0.1337) >>> BrE (0.0322) > NZE (0.0176) ≈ CanE (0.0154); for discourse markers (in red), the order and values of the variation coefficients are: BrE (0.0619) >> CanE (0.025) ≈ AusE (0.0227) > NZE (0.0104); for filled pauses (in green), we obtained BrE (0.0823) >> CanE (0.0447) > AusE (0.0325) > NZE (0.0128).

dependence scores for what corresponds to a three-way interaction, which is represented in Figure 6.

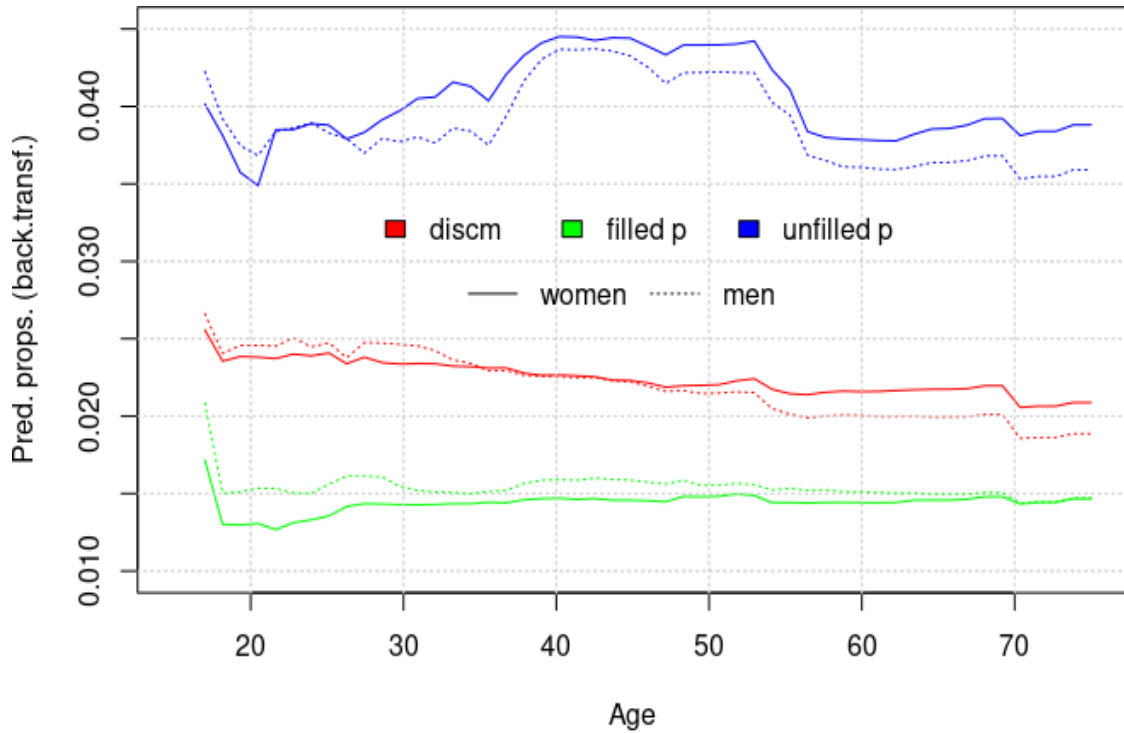


Figure 6: Predicted proportions of fluencemes for AGE:GENDER:FLUTYPE

The overall effect of FLUTYPE is the one we already know from above, and the overall effect of GENDER is indeed tiny, but there are at least some small tendential differences in how AGE is correlated with the response; for instance,

- unfilled pauses are most frequent for middle-aged speakers (both men and women);
- for discourse markers, there is a negative correlation with AGE, which might be ever so slightly stronger for men;
- filled pauses are most frequent for the youngest speakers, but largely remain uncorrelated with AGE once speakers reach 30 years of age.

4.4. The effect of EDUCATION_COARSE

The effect of EDUCATION_COARSE is weak and only slightly suggestive. Figure 7 appears to indicate that

- filled pauses are mildly negatively correlated with degree of education;

- unfilled pauses seem very weakly negatively correlated with degree of education;
- discourse markers are very weakly positively correlated with degree of education.

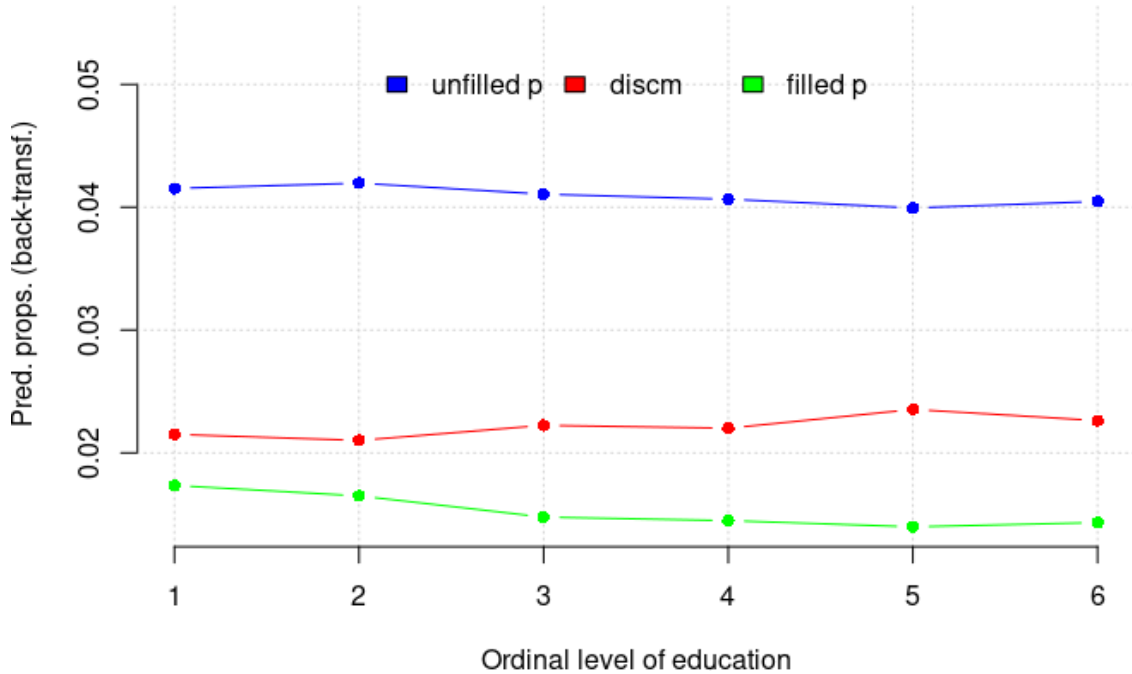


Figure 7: Predicted proportions of fluencemes for EDUCATION_COARSE

5. DISCUSSION, CONCLUSION AND OUTLOOK

This study explored the number and distribution of fluencemes across four L1-varieties of English, accounting for the influence of sociolinguistic variables. To our knowledge, it is the first study to investigate different fluencemes in combination across several L1-varieties of English. Looking at fluency in this holistic way enabled us to uncover complex and interactional effects between the different variables. We will discuss these findings in relation to our research questions in the following.

Regarding our first research question, that is, whether there are differences in fluenceme use between various English varieties, the answer is both affirmative and negative. While the variety of English indeed made a difference, the distribution of fluencemes across all four varieties remains similar. Notably, NZ English speakers generally use fewer fluencemes, particularly fewer unfilled pauses but, as we cautioned above, this could be due to the small number of NZ English speakers in our dataset or stem from potentially different transcription practices. Despite this, the overall

distribution of fluencemes is consistent across all varieties: unfilled pauses are used most frequently, followed by discourse markers, and then filled pauses. While it is reasonable to expect a similar degree of planning pressure among monolingual L1-English speakers, we had set out to test if individual varieties might exhibit unique preferences for specific types of fluencemes. However, the data suggest a striking similarity in terms of fluenceme frequencies, which might be due to their similar cognitive-functional role in speech planning. This points to an underlying “common [fluenceme] core” (Quirk *et al.* 1985: 16) across L1-English speakers. This core is

present in all varieties, so that, however esoteric a variety may be, it has running through it a set of [...] characteristics that are present in all the others. It is that fact that justifies the application of the name ‘English’ to all the varieties (Quirk *et al.* 1985: 16).

Given this comparable set of fluencemes across inner-circle English varieties, future studies should explore the specific realisations of fluencemes, that is, which types of discourse markers (e.g. *like* vs. *you know* vs. *I don’t know*) or which types of filled pauses (e.g. *uh* vs. *erm* vs. *ur*) along with potentially different phonological realisations are preferably used in certain varieties (e.g. Miller 2009; Tottie 2011, 2015; Fruehwald 2016), to determine if they qualify as *allo*-fluencemes. Such future investigations might yield more pronounced findings when examined in greater qualitative depth, such as by considering the position of the fluenceme within the utterance and the linguistic context following the fluenceme under scrutiny. Moreover, it would be particularly interesting to compare these findings with speakers of English as a second or foreign language. Understanding fluenceme preferences across different types of English varieties would significantly contribute to the discussion about a possible common fluenceme core.

As for the second research question —whether sociolinguistic variables influence the choice of particular fluencemes— the answer is only mildly affirmative. Contrary to our expectations based on previous research outlined in Section 2, the sociolinguistic variables we examined (gender, age, education, and occupation) exhibited much milder effects on speakers’ fluency than we had anticipated. This is particularly surprising in the case of age. Previous studies (e.g. Schow *et al.* 1978; Albert 1980) suggested that older individuals exhibit more planning phases than younger people, which we anticipated would correspond to a higher number of fluencemes in their speech. However, our findings did not strongly support this assertion, indicating that the influence of age on fluenceme use is subtler than previously thought. Instead, our findings revealed that the

highest number of unfilled pauses actually occurs in middle-aged speakers of both genders, while the number of filled pauses remains relatively stable from adolescence onwards for both genders. This result is puzzling, but the age-related effects, or lack thereof, are robust across varieties and a large number of speakers. Our findings thus seem to support the findings of a recent longitudinal study by Beier *et al.* (2023), which also found no significant increase in filled pauses among older individuals. Instead, their study suggested that older speakers tend to use different types of fluencemes, such as word repetitions, and exhibit different strategies to overcome planning phases, like slower speech. To obtain more reliable results regarding age-related fluency profiles, future studies should examine a broader range of fluenceme types across different English varieties. This would help clarify whether age-related patterns in fluency are consistent across various contexts and varieties.

The predicted age effect on discourse markers is also only mildly evident in our data, with their usage slightly decreasing as age increases. However, since our study did not distinguish between different types of discourse markers, this overall decline is relatively hard to interpret, since previous research indicates that some discourse markers increase with age, while others decrease or even remain constant (e.g. Miller 2009; Reichelt 2021). Given the vast variety of discourse markers, it seems necessary that future studies examine this variation more closely while controlling for sociolinguistic and geographic variables.

Turning to gender effects, gender also had an only mild and varied effect on the fluencemes we investigated. Whereas speakers from both genders show generally the same distributional patterns of fluenceme use, male speakers use more filled pauses than their female counterparts in all age groups (see also Shriberg 1996 or Bortfeld *et al.* 2001). Women use unfilled pauses more frequently than men from their early twenties onwards, a finding that we have not seen to have been discussed previously. As far as discourse markers are concerned, men use them more frequently until the age of approximately 35, when women's frequencies overtake those of men. These observations are only partly in line with Laserna *et al.*'s (2014) results, who find comparable frequencies of filled pauses across genders and ages, whereas discourse markers were more common among women in their dataset. Our findings add to the complexity of previous research output (e.g. Laserna *et al.* 2014; Scheuringer *et al.* 2017; Sokołowski *et al.* 2020). The gender effect on fluenceme use and its connection to utterance fluency therefore needs to be followed

up on further. One hypothesis in this vein has been put forward by Weiss *et al.* (2006), who suggest that men and women simply make use of different fluency-enhancing strategies to overcome planning phases, which at least partly also figures in our data.

However, although our findings do not clearly support specific gender- or age-related fluency profiles, we were still able to observe some overall trends, especially when looking at the three-way interaction between age, gender, and fluencemes. This emphasises the importance of controlling for such variables when investigating fluency. Potential follow-up studies could compare these findings with other fluenceme types to see if men/women or younger/older speakers make use of other strategies when they encounter planning phases (e.g. truncations, syllable-lengthening, repeats, etc.) or if generally different ways of easing the cognitive load of speaking can be observed, such as, for example, through a more frequent use of routinised formulae or even slower speech as in Götz's (2013) fluency groups or Dumont's (2018) (dis)fluency profiles.

Regarding the effect of speakers' educational level on their fluency, we observed only weak effects. There was a slight trend suggesting that more educated speakers use fewer filled and unfilled pauses (see also Stubbe and Holmes 1995), while an opposite, equally weak trend was observed for discourse markers. However, given the mild nature of these effects in our data, we do not consider ourselves in a position to draw any significant inferences from them at this point. Further research with a larger and more varied sample may be necessary to clarify the relationship between educational level and fluenceme use.

There are also some caveats to our study that we need to mention and which have the potential to turn into future research opportunities. For one, while our purely frequency-based approach to predicting fluencemes has revealed some interesting insights, the drawbacks have also become apparent. Especially for discourse markers and filled pauses, where we extracted a variety of different types, a purely frequency-based approach is extremely limited when it comes to interpreting the findings. It would therefore be particularly worthwhile to study the use of these different types along with their communicative functions and their utterance positions in a follow-up study in order to be able to reveal true variety-specific preferences of using certain discourse markers (in different functions or utterance positions) over others. On a more general level, this is especially relevant because at present our analysis does not include any linguistic and/or otherwise contextual level-1 predictors, that is, variables that describe context of an

individual choice for a particular fluenceme (see Gries 2024). For instance, it would be very interesting to determine whether there is a correlation between the (interaction of age and gender and) variety when it comes to investigating, for example which fluenceme is preferred depending on the complexity of the material directly following it.

Another phenomenon that falls into the same level-1 context category and that we have not been able to account for in the present study is the tendency of fluencemes to cluster together. For instance, discourse markers have been documented to frequently co-occur with filled or unfilled pauses or other discourse markers (e.g. Crible *et al.* 2017; Pons Bordería 2018). Therefore, looking at fluencemes in isolation only reveals limited information as for the number and quality of strategies speakers of different varieties of English require to fill planning phases. The same applies to utterance positions in which certain (clusters of) fluencemes occur, and even more so for the linguistic characteristics of a lexeme following a fluenceme (cluster), which we were not able to account for in the context of the present study. However, it would be extremely revealing to investigate in future studies.

Finally, we have to acknowledge that we have only investigated a limited number of only four inner-circle varieties of English. While this approach yielded promising results, there are contrastive research opportunities to take into consideration, such as how different variety types of English establish fluency, and whether the potential common fluenceme core can be extended to other types of Englishes.

REFERENCES

- Aijmer, Karin. 2002. *English Discourse Particles: Evidence from a Corpus*. Amsterdam: John Benjamins.
- Albert, Martin L. 1980. Language in normal and dementing elderly. In Lauraine K. Obler and Martin L. Albert eds. *Language and Communication in the Elderly*. Lexington, MA: DC Heath and Co, 145–150.
- Beeching, Kate. 2016. *Pragmatic Markers in British English: Meaning in Social Interaction*. Cambridge: Cambridge University Press.
- Beier, Eleonora J., Suphasiree Chantavarin and Fernanda Ferreira. 2023. Do disfluencies increase with age? Evidence from a sequential corpus study of disfluencies. *The Psychology of Aging* 38/3: 203–218.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.
- Börjars, Kersti and Kate Burridge. 2010. *Introducing English Grammar* (second edition). London: Hodder Education.

- Bortfeld, Heather, Silvia D. Leon, Jonathan E. Bloom, Michael F. Schober and Susan E. Brennan. 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role and gender. *Language and Speech* 44/2: 123–147.
- Brand, Christiane and Sandra Götz. 2011. Fluency versus accuracy in advanced spoken learner language: A multi-method approach. *International Journal of Corpus Linguistics* 16/2: 255–275.
- Brinton, Laurel J. 1996. *Pragmatic Markers in English: Grammaticalization and Discourse Functions*. Berlin: Mouton de Gruyter.
- Burke, Isabelle and Kate Burridge. 2023. From *a bit of processed cheese* to *a bit of a car accident* and *a little bit of 'oh really'* – The journey of Australian English *a bit (of)*. *Journal of Pragmatics* 209: 15–30.
- Burridge, Kate. 2014. *Cos* – A new discourse marker for Australian English? *Australian Journal of Linguistics* 34/4: 524–548.
- Burridge, Kate and Margaret Florey. 2002. 'Yeah-no He's a Good Kid': A discourse analysis of *Yeah-no* in Australian English. *Australian Journal of Linguistics* 22/2: 149–171.
- Butcher, Alan. 1980. Pause and syntactic structure. In Hans-Werner Dechert and Manfred Raupach eds. *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler*. The Hague: Mouton de Gruyter, 85–90.
- Buyse, Lieven and Meghan Blanchard. 2022. L1 and non-L1 perceptions of discourse markers in English. *Pragmatics & Cognition* 29/2: 222–245.
- Chafe, Wallace L. ed. 1980. *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Norwood, NJ: Ablex.
- Chambers, Francine. 1997. What do we mean by fluency? *System* 25/4: 535–544.
- Clark, Herbert. H. and Jean E. Fox Tree. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition* 84: 73–111.
- Corley, Martin, Lucy J. MacGregor and David I. Donaldson. 2007. It's the way that you, *er*, say it: Hesitations in speech affect language comprehension. *Cognition* 105/3: 658–668.
- Crible, Ludivine. 2018. *Discourse Markers and (Dis)fluency: Forms and Functions across Languages and Registers*. Amsterdam: John Benjamins.
- Crible, Ludivine, Lisbeth Degand and Gaëtanelle Gilquin. 2017. The clustering of discourse markers and filled pauses: A corpus-based French-English study of (dis)fluency. *Languages in Contrast* 17/1: 69–95.
- Drommel, Raimund H. 1980. Towards a subcategorization of speech pauses. In Hans-Werner Dechert and Manfred Raupach eds. *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler*. The Hague: Mouton, 227–238.
- Dumont, Amandine. 2018. *Fluency and Disfluency. A Corpus Study of Non-native and Native Speakers (Dis)fluency Profiles*. PhD Thesis: Université Catholique de Louvain.
- Foster, Pauline and Parvaneh Tavakoli. 2009. Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language Learning* 59/4: 866–896.
- Fox, Jon and Sanford Weisberg. 2019. *An R Companion to Applied Regression* (third edition). Thousand Oaks CA: Sage.
- Fruehwald, Josef. 2016. Filled pause choice as a sociolinguistic variable. In Helen Jeoung ed. *University of Pennsylvania Working Papers in Linguistics: Selected Papers from New Ways of Analyzing Variation (NWAV) 22* (second edition). Pennsylvania: Penn Graduate Linguistics Society, 41–49.
<https://repository.upenn.edu/handle/20.500.14332/45126> (2 January, 2025.)

- Fuller, Janet M. 2003. Discourse marker use across speech contexts: A comparison of native and non-native speaker performance. *Multilingua* 22: 185–208.
- Fung, Loretta and Ronald Carter. 2007. Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics* 28/3: 410–439.
- Ganzeboom, Harry B.G. 2010. A new International Socio-Economic Index (ISEI) of occupational status for the International Standard Classification of Occupation 2008 (ISCO-08) constructed with data from the ISSP 2002–2007. Paper presented at *Annual Conference of International Social Survey Programme*, Lisbon, May 1 2010.
- Goldman-Eisler, Frieda. 1961. A comparative study of two hesitation phenomena. *Language and Speech* 4/1: 18–26.
- Götz, Sandra. 2013. *Fluency in Native and Nonnative English Speech*. Amsterdam: John Benjamins.
- Grant, Lynn E. 2010. A corpus comparison of the use of *I don't know* by British and New Zealand speakers. *Journal of Pragmatics* 42/8: 2282–2296.
- Greenbaum, Sidney and Gerald Nelson. 1996. The International Corpus of English (ICE) Project. *World Englishes* 15/1: 3–15.
- Greenwell, Brandon M. 2017. pdp: An R package for constructing partial dependence plots. *The R Journal* 9/1: 421–436.
- Gries, Stefan Th. 2020. On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory* 16/3: 617–647.
- Gries, Stefan Th. 2021. *Statistics for Linguistics with R* (third edition). Berlin: Mouton de Gruyter.
- Gries, Stefan Th. 2024. Against level-3-only analyses in corpus linguistics. *ICAME Journal* 48/1: 1–25.
- Hasselgren, Angela. 2002. Learner corpora and language testing: Smallwords as markers of learner fluency. In Sylviane Granger, Joseph Hung and Stephanie Petch-Tyson eds. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins, 143–173.
- Kachru, Brach. B. 1985. Standards, codification and sociolinguistic realism: The English language in the outer circle. In Randolph Quirk and Henry G. Widdowson eds. *English in the World: Teaching and Learning the Language and Literatures*. Cambridge: Cambridge University Press, 11–30.
- Kirjavainen, Minna, Ludivine Crible and Kate Beeching. 2022. Can filled pauses be represented as linguistic items? Investigating the effect of exposure on the perception and production of *um*. *Language and Speech* 65/2: 263–289.
- Kjellmer, Göran. 2003. Hesitation. In defence of ER and ERM. *English Studies* 84/2: 170–198.
- Laserna, Charlyn M., Yi-Tai Seih and James W. Pennebaker. 2014. *Um ... who like says you know*: Filler word use as a function of age, gender, and personality. *Journal of Language and Social Psychology* 33/3: 328–338.
- Leuckert, Sven and Sofia Rüdiger eds. 2021. *World Englishes 40/4: Special Issue on Discourse Markers and World Englishes*.
- Lounsbury, Floyd G. 1954. Transitional probability, linguistic structure and systems of habit-family hierarchies. In Charles E. Osgood and Thomas A. Sebeok eds. *Psycholinguistics: A Survey of Theory and Research Problems*. Bloomington, IN: Indiana University Press, 93–101.
- Loy, Jia E., Hannah Rohde and Martin Corley. 2017. Effects of disfluency in online interpretation of deception. *Cognitive Science* 41/6: 1434–1456.

- Maclay, Howard and Charles E. Osgood. 1959. Hesitation phenomena in spontaneous English speech. *WORD: Journal of the International Linguistics Association* 15: 19–44.
- Manhire, Laura. 2021. *Yeah nah she'll be right: An attitudinal study of 'yeah nah' in New Zealand English*. MA Thesis. Canterbury: University of Canterbury dissertation.
- Miller, Jim. 2009. *Like* and other discourse markers. In Pam Peters, Peter Collins and Adam Smith eds. *Comparative Studies in Australian and New Zealand English*. Amsterdam: John Benjamins, 315–336.
- O'Connell, Daniel C. and Sabine Kowal. 2005. *Uh* and *Um* revisited: Are they interjections for signaling delay? *Journal of Psycholinguistic Research* 34/6: 555–576.
- OECD. 1999. *Classifying Educational Programmes: Manual for ISCED-97 Implementation in OECD Countries*, OECD Publishing, Paris, <http://www.oecd.org/education/1841854.pdf> (23 May, 2024.)
- Osborne, John. 2013. Fluency, complexity and informativeness in native and non-native speech. In Gaëtanelle Gilquin and Sylvie De Cock eds. *Errors and Disfluencies in Spoken Corpora*. Amsterdam: John Benjamins, 140–161.
- Pawley, Andrew and Frances H. Syder. 1983. Two puzzles for linguistic theory: Native-like selection and native-like fluency. In Jack Richards and Richard W. Schmidt eds. *Language and Communication*. London: Longman, 191–226.
- Pons Bordería, Salvador. 2018. The combination of discourse markers in spontaneous conversations: Keys to undo a gordian knot. *Revue Romane* 53/1: 121–158.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
- Reichelt, Susan. 2021. Recent developments of the pragmatic markers *kind of* and *sort of* in spoken British English. *English Language and Linguistics* 25/3: 563–580.
- Revis, Melanie and Tobias Bernaisch. 2020. The pragmatic nativisation of pauses in Asian Englishes. *World Englishes* 39/1: 135–153.
- Riggenbach, Heidi. 1991. Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes* 14/4: 423–441.
- Scheuringer, Andrea, Ramona Wittig and Belinda Pletzer. 2017. Sex differences in verbal fluency: The role of strategies and instructions. *Cognitive Processing* 18/4: 407–417.
- Schow, Ronald L., John M. Christensen, John M. Hutchinson and Michael Nerbonne. 1978. *Communication Disorders of the Aged: A Guide for Health Professionals*. Baltimore, MD: University Park Press.
- Schweinberger, Martin. 2018. The discourse particle *eh* in New Zealand English. *Australian Journal of Linguistics* 38/3: 395–420.
- Segalowitz, Norman. 2010. *Cognitive Bases of Second-Language Fluency*. London: Routledge.
- Shriberg, Elizabeth. 1996. Disfluencies in switchboard. *Proceedings of the International Conference on Spoken Language Processing (ICSLP '96)*, Volume addendum, 11–14. Philadelphia, PA, 3–6 October. <http://www.asel.udel.edu/icslp/cdrom/vol3/1031/a1031.pdf> (2 January, 2025.)
- Sokołowski, Andrzej, Ernest Tyburski, Anna Sołtys and Ewa Karabanowicz. 2020. Sex differences in verbal fluency among young adults. *Advances in Cognitive Psychology* 16/2: 92–102.
- Strobl, Carolin, James D. Malley and Gerhard Tutz. 2009. An introduction to recursive partitioning: Rationale, application and characteristics of classification and

- regression trees, bagging and random forests. *Psychological Methods* 14/4: 323–348.
- Stubbe, Maria and Janet Holmes. 1995. *You know, eh* and other ‘exasperating expressions’: An analysis of social and stylistic variation in the use of pragmatic devices in a sample of New Zealand English. *Language & Communication* 15/1: 63–88.
- Svartvik, Jan. 1990. The TESS project. In Jan Svartvik ed. *The London Lund Corpus of Spoken English: Description and Research*. Amsterdam: Rodopi, 203–218.
- Tagliamonte, Sali. 2005. *So who? Like how? Just what?* Discourse markers in the conversations of young Canadians. *Journal of Pragmatics* 37/11: 1896–1915.
- Tottie, Gunnel. 2011. *Uh* and *Um* as sociolinguistic markers in British English. *International Journal of Corpus Linguistics* 16/2: 173–197.
- Tottie, Gunnel. 2015. *Uh* and *um* in British and American English: Are they words? Evidence from co-occurrence with pauses. In Rena Torres Cacoullos, Nathalie Dion and André Lapierre eds. *Linguistic Variation: Confronting Fact and Theory*. New York: Routledge, 38–55.
- UNESCO. 2006. International Standard Classification of Education (ISCED 1997). Paris: UNESCO. https://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-1997-en_0.pdf (2 January, 2025.)
- Van Gelderen, Elly. 2010. *An Introduction to the Grammar of English: Revised Edition*. Amsterdam: John Benjamins.
- Vine, Bernadette. 2016. Pragmatic markers at work in New Zealand. In Lucy Pickering, Eric Friginal and Shelley Staples eds. *Talking at Work: Communicating in Professions and Organizations*. London: Palgrave Macmillan, 1–25.
- Weiss, Elisabeth M., Daniel Ragland, Colleen M. Brensinger, Warren B. Bilker, Eberhard A. Deisenhammer and Margarete Delazer. 2006. Sex differences in clustering and switching in verbal fluency tasks. *Journal of the International Neuropsychological Society* 12/4: 502–509.
- Wolk, Christoph, Sandra Götz and Katja Jäschke. 2021. Possibilities and drawbacks of using an online application for semi-automatic corpus analysis to investigate discourse markers and alternative fluency variables. *Corpus Pragmatics* 5/1: 7–36.
- Wright, Marvin N. and Andreas Ziegler. 2017. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* 77/1: 1–17.
- Wright, Marvin N., Andreas Ziegler and Inke R. König. 2016. Do little interactions get lost in dark random forests? *BMC Bioinformatics* 17: article 145. <https://doi.org/10.1186/s12859-016-0995-8> (2 January, 2025.)

Corresponding author

Sandra Götz
Philipps University Marburg
Department of English and American Studies
Wilhelm-Röpke-Str. 6
35032 Marburg
e-mail: goetz-lehmann@uni-marburg.de

received: November 2024
accepted: January 2025

The *Construction Complexity Calculator* (ConPlex): A tool for calculating Nelson's (2024) construction-based complexity measure

Christopher R. Cooper
Rikkyo University / Japan

Abstract – The current study aims to increase the accessibility of Nelson's (2024) recently suggested construction-based complexity measure by providing a tool that can calculate the measure for single or multiple texts. To validate the tool, complexity scores for the *International Corpus Network of Asian Learners of English* corpus (ICNALE) were compared with Nelson's (2024) results. In addition, complexity scores were calculated for a new dataset, the *Common European Framework of Reference English Listening Corpus* (CEFR), along with the *MERLIN* corpus, which includes learner writing samples from learners of Czech, German, and Italian. Complexity scores generally increased across CEFR levels in all of the datasets. However, the complexity scores in the current study tend to be higher than the original study due to differences in the sentence splitting approach. The sentence tokenisation method used is deemed to be more appropriate, and it may be concluded that the *Construction Complexity Calculator* (ConPlex) tool accurately calculates Nelson's measure. It is hoped that the tool will allow researchers to calculate the complexity of constructions at the text level for a wide range of research purposes.

Keywords – constructions; complexity; corpus tool creation; corpus tool validation; language development

1. INTRODUCTION

For open science to live up to its name, the code used for data analysis should be shared, and researchers should be actively involved in the development of freely available tools (Mizumoto 2024). Several tools are already available for measuring an array of indices that can be utilised to assess complexity and other elements of the production of second language (L2) users and other texts. These include *Coh-Metrix* (McNamara *et al.* 2014), the tool for the *Automatic analysis of Syntactic Sophistication and Complexity* (TAASSC; Kyle 2016), and the tool for the *Automatic Analysis of Lexical Diversity* (Kyle *et al.* 2021). The underlying code has also been made available for the latter two



tools on GitHub.¹ The present study is a further contribution to open science and aims to 1) design a tool —the *Construction Complexity Calculator* (ConPlex)— to calculate a recently suggested construction-based complexity measure by Nelson (2024) and 2) to investigate the applicability of the tool to various types of texts in multiple languages, namely English, Czech, German, and Italian. The paper does not only describe the development of the tool in detail, along with its validation by comparing complexity scores output by the tool with some of Nelson’s (2024) results but also highlights what researchers should be aware of when using the tool in their research.

The paper is organised as follows. Section 2 discusses the notion of ‘complexity’ in linguistics. Section 3 offers information on the corpora used to validate the accuracy of ConPlex and describes how the tool has been produced. Section 4 deals with the tool validation results as well as its potential uses and limitations. Finally, Section 5 offers some concluding remarks.

2. CONSTRUCTION-BASED COMPLEXITY

There has been some recent debate about the types of measure that should be used to represent the construct of ‘complexity’ in linguistics or second language acquisition (SLA), and there has been no consensus about a common measure so far (Ehret *et al.* 2023: 2). In their theoretical and methodological overview, Bulté *et al.* (2024) assert that there should be a more restricted interpretation of complexity and suggest a set of core measures that should be used to increase replicability and knowledge accumulation. They put forward a list of eight core measures of complexity which include moving-average type-token ratio (MATTR) for several indices, and various ratio-based measures at the word, phrase, clause, T-unit, and AS-unit level. Bulté *et al.*’s (2024) manuscript has provoked several ‘open peer commentary’ responses in the same journal. For instance, the response by Biber *et al.* (2024: 1–2) points out that the ‘omnibus’ measures suggested by Bulté *et al.* (2024) disregard the syntactic functions of grammatical structures. They liken this to a biologist taking a reductionist method and operationalising the complexity of forests by simply calculating the average height of trees and the mean number of branches per tree. Biber *et al.*’s comment seems a valid point, as fine-grained measures can reveal intricate details about the grammatical

¹ <https://github.com/kristopherkyle>

complexity of a language. However, there is also a place for omnibus measures that represent the interaction of multiple features. If omnibus methods that match a theoretical construct in the field are selected, they may have the benefit of being applied to a wider range of texts. Omnibus measures could potentially be applied to multiple registers or languages without the need to create a taxonomy of grammatical possibilities in the target domain. The most appropriate measures might not necessarily be the ones suggested by Bulté *et al.* (2024). Appropriate measures can be selected to match the theoretical beliefs of the researcher and to answer specific research questions. There has also been disagreement regarding the use of the sentence as a syntactic unit. On the one hand, Bulté *et al.* (2024) claim that there is no linguistic definition of the sentence that is agreed upon and suggest that it is an unusable unit for oral data or for analysing texts produced by writers whose punctuation skills are limited. On the other hand, Lu (2024) points out that the sentence as a unit is intuitively useful in writing.

Nelson (2024) takes an alternative approach to the measuring of complexity that is grounded in Complexity Theory and Construction Grammar. In some ways, Complexity Theory is also in line with the abovementioned measures, as “the behaviour of complex systems emerges from the interactions of its components” (Larsen-Freeman 1997: 143), and it is not concentrated on a specific component. Rather, complexity theorists are interested in how the parts of a complex system interact (Larsen-Freeman 1997), not merely in the production of a vast taxonomy of individual factors (Larsen-Freeman and Cameron 2008: 206). More recently, Larsen-Freeman (2017) has described Complexity Theory as a metatheory which also requires a theory of language and how it develops. One of these metatheories is Construction Grammar, in which Goldberg (2003: 219) defines constructions as “stored pairings of form and function, including morphemes, words, idioms, partially lexically filled and fully general linguistic patterns” and further argues that

any linguistic pattern is recognised as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognised to exist.

Construction Grammar differs from other grammar descriptions in that it aims to account for the whole of the language. However, no finite typology of all of the possible constructions in the English language has been agreed upon. Therefore, Nelson’s (2024: 13) measure seeks to account for “how the diversity of constructions used impacts the

statistical properties of the texts a person produces.” The measure is calculated as shown below:

$$C(d) = \frac{1}{N} \sum_{i=1}^N D(S_i) P(S_i)$$

The complexity of the document $C(d)$ is calculated as the mean diversity $D(S_i)$ and the mean productivity $P(S_i)$ of all of the sentences in the document. First the text is part-of-speech (POS) tagged, then the diversity of each sentence is calculated by partitioning the tags in the sentence into lists of tag pairs. A list of tags is taken and partitioned into pairs at an overlap of one, meaning that the last tag in the pair (T1, T2) is the first in the succeeding pair (T2, T3). For example, the sentence from Mary Shelley’s *Frankenstein* would be converted into tag pairs as shown in (1):

(1) **Sentence:** “There is something at work in my soul, which I do not understand.”

Tagged sentence: There_EX is_VBZ something_NN at_IN work_NN in_IN my_PRP\$ soul_NN which_WDT I_PRP do_VBP not_RB understand_VB

Tag pairs: (‘EX’, ‘VBZ’), (‘VBZ’, ‘NN’), (‘NN’, ‘IN’), (‘IN’, ‘NN’), (‘NN’, ‘IN’), (‘IN’, ‘PRP\$’), (‘PRP\$’, ‘NN’), (‘NN’, ‘ WDT ’), (‘WDT’, ‘PRP’), (‘PRP’, ‘VBP’), (‘VBP’, ‘RB’), (‘RB’, ‘VB’)

Next, the ‘Shannon entropy’ (Shannon 1948) of tag pairs is calculated, the mean of which contributes to the complexity score for the text. The productivity of each sentence is calculated as the entropy of word tag pairings minus the entropy of tags. 1 is added to the productivity calculation to account for situations when the entropy of word tag pairings is 0 or less than 1. The sentence in the *Frankenstein* example above is comprised of 13 unique pairings of a tag and a word but only ten tags. This is because there are three nouns (*something*, *work*, and *soul*) tagged as ‘NN’ and two prepositions (*at* and *in*) tagged as ‘IN’. This information is used in the productivity calculation. Given pairs of words and their tags (e.g., pairs = {{there, EX}, {is, VBZ}, ... {understand, VB}}) which can be represented as two ordered lists (i.e., tags = {EX, VBZ, ... VB} and words = {there, is, ...understand}), productivity is calculated as the conditional entropy of the words given the tags or $H(\text{words} \mid \text{tags}) = H(\text{tags}, \text{words}) - H(\text{tags})$. The complexity of the sentence is calculated by multiplying diversity and productivity, as they are held to interact. The complexity of a document is taken as the mean of the complexity of all the sentence-level complexity scores in the document.

Nelson (2024) clearly shows how measuring diversity in this way can represent the complexity of constructions using POS graphs. In addition, the sentence in (2), taken from F. Scott Fitzgerald's *The Great Gatsby*, is used as an example to illustrate the need for the productivity element of the measure.

- (2) The apartment was on the top floor - a small living-room, a small dining-room, a small bedroom, and a bath.

If the first two occurrences of *small* in the sentence were replaced by *cosy* and *spacious*, the productivity score for the sentence would increase due to the wider range of word tag pairings. This, in turn, would increase the complexity score. Objectively, the sentence containing a wider variety of adjectives would likely be considered more complex by most if not all readers. Calculating complexity in this way is in line with one of Larsen-Freeman and Cameron's (2008: 206) methodological principles to identify collective variables that are characteristic of multiple elements interacting within a system. Although only one complexity score is output for each text, the score represents the interaction of components within the system, as opposed to a vast taxonomy of individual scores that represent individual components in the system. In this sense, the measure could be said to be more in line with Complexity Theory.

In addition to proposing the measure, Nelson (2024) also applies it to several datasets and shows that complexity scores increase 0.015 per month with L1 acquisition data from children (MacWhinney 2000). Furthermore, results from a Bayesian hierarchical model show that an increase in complexity measure scores correlates with the proficiency level of L2 learners in data taken from the *International Corpus Network of Asian Learners of English* (ICNALE; Ishikawa 2023). The measure also shows strong correlations with traditional readability measures. Moreover, when comparing U.S. presidential campaign speeches from 2016, results from a mixed effects model suggest that complexity is not affected by text length.

Although the theory behind Nelson's (2024) construction-based complexity measure has been summarised here, it is highly recommended that readers interested in using the ConPlex engage with Nelson's paper, where a more detailed theoretical background is provided.

3. TOOL CREATION

This section introduces the corpora used to validate the accuracy of ConPlex and describes the technical steps taken to produce the tool.

3.1. Corpora used to validate the tool

To assess the accuracy of the tool output, the complexity scores from the spoken monologues and written essays in ICNALE were compared with Nelson's (2024) results, which were obtained after contacting the researcher. ICNALE includes 4,400 spoken monologues and 5,600 written essays produced by university students in ten countries across Asia. As such, it is one of the largest publicly available learner corpora and includes texts at the A2, B1_1, B1_2, and B2+ CEFR levels. The corpus also includes data produced by native speakers of English who completed the same spoken and written tasks as the L2 users. The transcripts of the monologues and written essays were used in the analysis and no further pre-processing was done to the texts.

To evaluate the tool further, complexity scores were calculated for a new dataset, the *CEFR English Listening Corpus*. This corpus was compiled by the author from listening texts that are freely available for language study online. The first source of texts includes two British Council websites² that feature listening texts and videos that have been produced for the website, and *YouTube* videos that are not produced by the British Council. Each text has been assigned a CEFR level by the producers of the website. In some cases, the CEFR level is broad, spanning several levels, such as B1-B2 and B2-C1-C2. In these cases, the lowest CEFR level was counted.

In order to increase the size of the corpus, listening texts from the CEFR-aligned Cambridge exams,³ which are available online for exam preparation, were added. Although these texts have a different purpose, they were selected for the corpus to include a range of texts aimed at L2 learners of English for practicing and assessing their own listening. The final corpus size was 728 texts and 345,104 words, as can be noticed in Table 1, where more detailed information about the number and length of texts from each source is provided.

² <https://learnenglish.britishcouncil.org/> and <https://learnenglishteens.britishcouncil.org/>

³ <https://www.cambridgeenglish.org/learning-english/exam-preparation/>

Text source	Texts	Tokens	Text length (tokens)			
			<i>M</i>	<i>SD</i>	Min	Max
British Council	563	303,959	540	560	54	4,751
Cambridge Exams	165	41,145	249	200	47	851
Total	728	345,104	474	516	47	4,751

Table 1: Size of the *CEFR English Listening Corpus*

The distribution of texts by CEFR level and text type is shown in Table 2. A limitation of the *CEFR English Listening Corpus* is how imbalanced the dataset is. In particular, only six texts in the corpus are classified as C2 level, all of which were Cambridge listening texts. As the purpose of this part of the study was to investigate whether complexity scores increase across CEFR levels, the C2 level was still included as a separate class, instead of merging it with the C1 level.

Text type	A1	A2	B1	B2	C1	C2	Total
BC Listening	24	41	34	52	22	0	173
BC Videos	15	0	104	15	9	0	143
BC YouTube	0	2	34	157	54	0	247
Cambridge	18	32	38	53	18	6	165
Total	57	75	210	277	103	6	728

Table 2: Text types by CEFR level in the *CEFR English Listening Corpus*

In both cases the transcripts produced by the material creators were used. The British Council texts each had a transcript available online, which was presumably produced or at least checked by the creators of the website. For the Cambridge listening exams, transcripts were included in PDF files that were available with the audio files. The *CEFR English Listening Corpus* has not been released due to copyright.

While the desktop version of ConPlex currently supports only English texts, the underlying code is available as a *Python* notebook on GitHub.⁴ This enables researchers to adapt the code to investigate complexity in other languages. In the current study, the multilingual *MERLIN* corpus (Boyd *et al.* 2014) was used to evaluate the application of the complexity measure to three different languages. The *MERLIN* corpus features texts written by learners of Czech, German, and Italian that were taken from CEFR-aligned written examinations. CEFR levels containing fewer than ten texts per language were not included in the analysis. The final corpus size is described in terms of texts in Table 3, and in terms of tokens in Table 4. As part of the *MERLIN* corpus project, the CEFR level of each text was re-rated by specially trained testers in what was termed as a ‘fair

⁴ <https://github.com/cooperchris17/ConPlex> (accessed 10 March 2025).

rating’. This rating was used in the current study to represent the CEFR level. The plain text versions of the texts were used and no further pre-processing steps were taken.

Language	A1	A2	A2+	B1	B1+	B2	B2+	C1	Total
Czech		76	112	90	75	72			425
German	57	199	107	217	115	219	73	42	1,029
Italian	29	289	92	341	53				804

Table 3: Number of texts per CEFR level in the *MERLIN* corpus used in the current study

Language	Texts	Tokens	Text length (tokens)			
			<i>M</i>	<i>SD</i>	Min	Max
Czech	425	61,013	61,013	61,013	61,013	61,013
German	1,029	126,468	126,468	126,468	126,468	126,468
Italian	804	93,292	93,292	93,292	93,292	93,292
Total	2,258	280,773	474	280,773	280,773	280,773

Table 4: Size of the *MERLIN* corpus used in the current study

It should be noted that a complexity measure that has been designed and validated on the English language will not necessarily behave in the same way when used with other languages. Previous research has shown that there are several differences between the three languages in the *MERLIN* corpus when compared to English. For example, Kettunen (2014) measured the complexity of the European Union constitution written in 21 languages using a morphological complexity measure and two type-token ratio (TTR) metrics. Of the four languages considered in the current study, English and Italian were the least complex for all the measures, German was the most morphologically complex, and Czech had the highest TTR scores. It has also been pointed out that Czech has an overt inflectional morphology, whereas the morphological features of English are predominantly analytic (Hledíková and Ševčíková 2024). Also, when compared with German, the distance between form and meaning is often greater in English (Hawkins 2015). Due to these and other differences, complexity scores will not necessarily be comparable between languages. However, as the grammatical constraints of an individual language impact texts written or spoken in that language in a similar way, the measure should be suitable for at least an initial exploratory investigation of languages other than English.

3.2 Producing the tool

The tool was designed to be accessible to as many researchers as possible. While the code to calculate complexity in Nelson’s (2024) paper was written in *Mathematica*, *Python* was chosen for ConPlex. *Python* is one of the most widely used programming languages: it is highly readable and often used for natural language processing tasks. This makes it a suitable choice, as researchers may wish to adapt the code that is released with ConPlex. In addition, while Nelson (2024) worked with POS tagged texts that had been processed before the complexity analysis, ConPlex was designed to accept plain text files and handle POS tagging within the tool. This is simpler for the end user, as they do not need to use a separate tool to tag their texts. Furthermore, it avoids the problem of dealing with output from different taggers, which are likely to follow inconsistent formatting standards. *Stanza* (Qi *et al.* 2020) is used in ConPlex to split the texts into sentences and to tag the texts with treebank-specific POS tags. Next, tokens tagged with the universal POS tag ‘PUNCT’, meaning all punctuation, are excluded from the analysis. Then, all words are converted to lower case to avoid words that are used more than once in the same sentence being counted as different words when they occur at the beginning of a sentence. *Stanza* was initially chosen to replicate the fact that Nelson (2024) used the *Stanford Tagger* in his analysis.⁵ The NLP library *SpaCy* (Honnibal *et al.* 2023) was also trialled, but inspection of the output showed that *Stanza* was more accurate for sentence tokenisation. Diversity, production, and complexity were calculated as described in Section 2 and the entropy function in *SciPy* (Virtanen *et al.* 2020) was used in the calculations. After trialling the code in a *Python* notebook, a downloadable app was created using *PyQt5*.⁶ Producing a downloadable app allows researchers who are not familiar with *Python* to use the tool with a graphical user interface.⁷ While it is assumed that most researchers will use the downloadable app for the analysis of English texts, a notebook with the *Python* code behind the tool is also available on GitHub. Widgets have been added to the notebook so the functionality is the same as the downloadable tool. Sharing the *Python* code in this way makes it possible for researchers to adapt the code to suit their research goals. Reasons for

⁵ <https://techfinder.stanford.edu/technology/stanford-part-speech-tagger>

⁶ <https://www.riverbankcomputing.com/software/pyqt/>

⁷ The *Windows* version of the app is available at <https://drive.google.com/file/d/1PljHorFOaXYTIar527GMaDibNAzk6Coo/view> (accessed 10 March 2025) and links to the *OSX* and *Linux* versions will be added to the app’s GitHub page at <https://github.com/cooperchris17/ConPlex> (accessed 10 March 2025) when they are ready.

adapting the code might include trying the complexity measure with other languages, as demonstrated in the current study, or changing the tagger to one that is more appropriate for the users' texts. The algorithm used in ConPlex is illustrated in Figure 1.

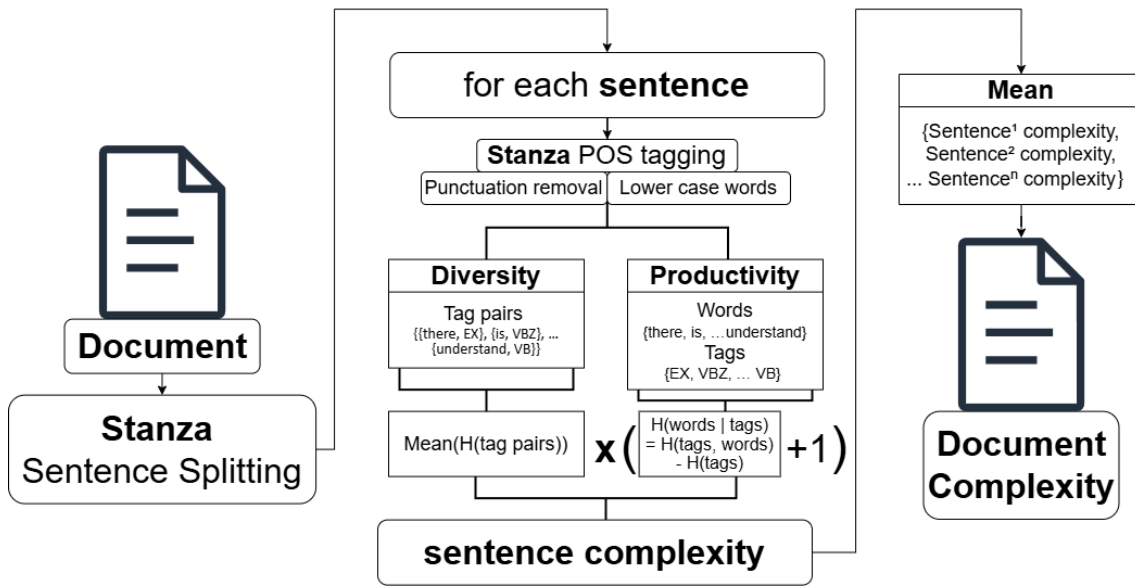


Figure 1: Visualisation of the algorithm used to calculate complexity in ConPlex

4. THE FINISHED TOOL

A screenshot of the downloadable app is shown in Figure 2. The tool has two input methods. The first option is to copy and paste one text into the textbox in the tool's interface, then click 'Process Text Input' to begin processing. For uploading one or multiple plain text files, the user can click 'Upload and Process Files'. For this option, processing begins as soon as the files have been selected. The tool outputs the mean complexity score for each text, along with the mean diversity and mean productivity scores. It is expected that most researchers will only use the complexity scores.

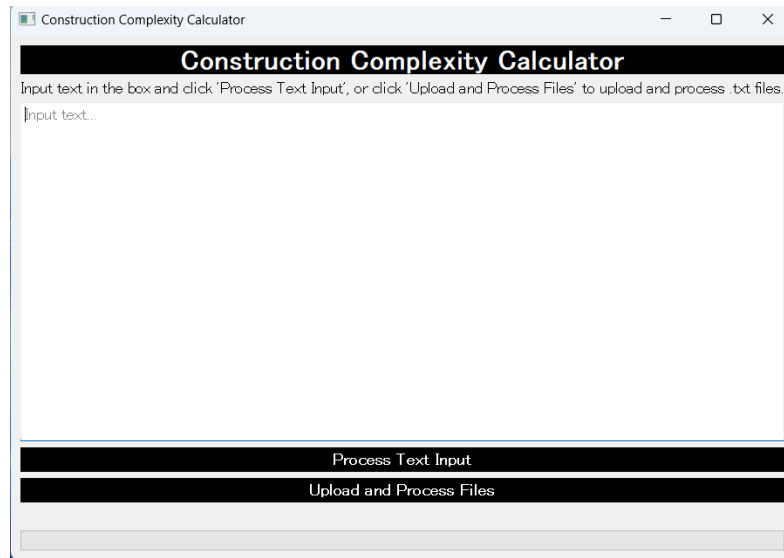


Figure 2: Screenshot of ConPlex in Windows

4.1. Tool validation results

The complexity scores calculated for the ICNALE texts are shown in Figure 3. The first point to note is that there is a progression across CEFR levels in both the present study and Nelson (2024). This is clearly evident in the central tendencies indicated by the boxplots and distributions shown in the violin plots. The progression cannot be described as linear, but this is in line with Complexity Theory, as “learning is not climbing a developmental ladder; it is not unidirectional.” (Larsen-Freeman 2017: 27). The two additional points to note about Figure 3 are 1) the difference between scores in the current study and Nelson (2024) and 2) the substantial number of outliers.

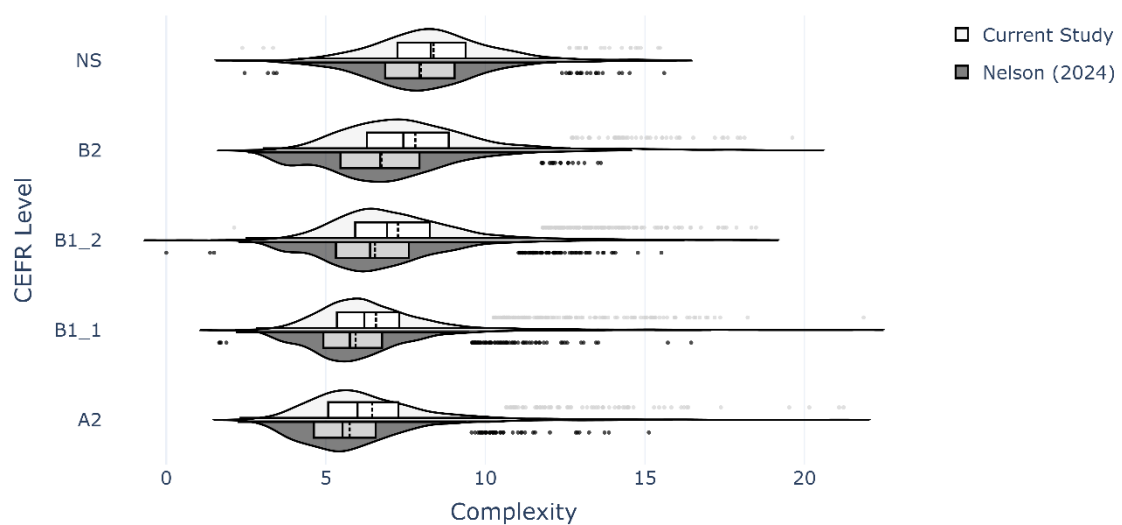


Figure 3: ICNALE complexity score distribution in the present study and Nelson (2024)

For a more fine-grained analysis, the difference between the scores for each text was calculated. Descriptive statistics are shown in Table 5 and the distribution of differences in complexity scores is illustrated in Figure 4. Positive values indicate higher scores in the current study and negative values indicate higher scores in Nelson (2024). In the current study, productivity scores are generally similar but slightly higher, and the diversity scores from Nelson (2024) are generally higher. The complexity scores are also slightly higher for the most part here. Some of the differences are extreme, with the maximum difference being 17.77. Despite these differences, Figure 4 illustrates that the differences for the majority of the texts are close to zero.

Measure	M	SD	Min	Q1	Mdn	Q3	Max
Complexity	0.70	1.99	-5.10	0.09	0.28	0.59	17.77
Diversity	-0.03	0.54	-2.14	-0.16	-0.09	-0.04	3.24
Productivity	0.14	0.27	-0.66	0.03	0.08	0.15	2.31

Table 5: The difference between individual texts in the current study and Nelson (2024)

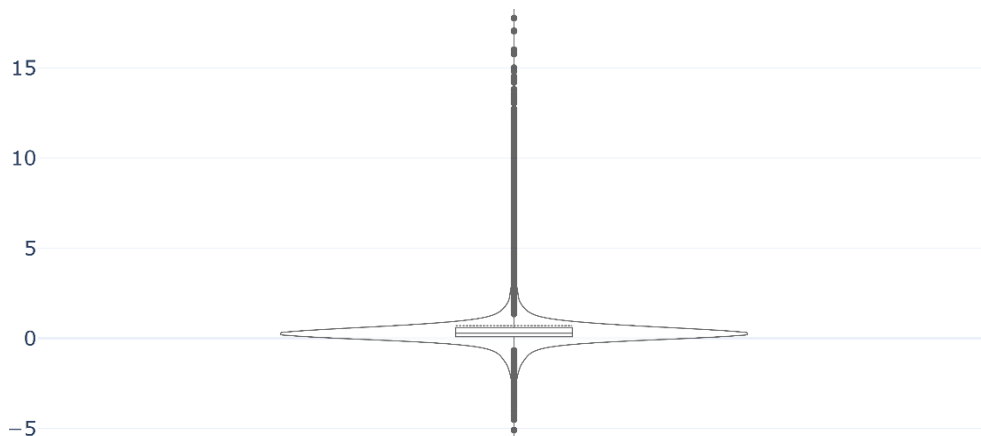


Figure 4: Difference in ICNALE complexity scores in the current study and Nelson (2024)

The reason for the difference in complexity scores was investigated by consulting texts that had large differences in the two studies. It was found that the differences seemed to be largely caused by sentence tokenisation. In the current study, texts were split into sentences using *Stanza*. However, Nelson (2024) used the period POS tag in the *Stanford Tagger* to split sentences at the following punctuation marks: “.”, “?”, and “!”. For example, the sentence “3. must fast show to arrive at the attention in serve.” was split into one sentence in this study, but in two in Nelson (2024). In addition, Nelson (2024) attempted to deal with learner texts that did not include accurate sentence

punctuation by splitting texts that featured less than two period POS tags into sentences every ten words using a bespoke function labelled ‘safeBreaks’. When an equivalent function to safeBreaks was added to the ConPlex code, the mean difference in scores reduced slightly to 0.51 ($SD = 1.42$). However, there were still extreme differences ($max = 13.25$) and the median difference was the same ($Mdn = 0.28$). There are some other small differences between the two implementations. The first is the handling of sentence internal punctuation: while ConPlex removes all punctuation based on the universal POS tag ‘PUNCT’ assigned by *Stanza*, Nelson’s (2024) code only seems to remove commas. In addition, there may be differences in the way that entropy is calculated in *Python* when compared with *Mathematica*.

Despite these differences, no changes were made to ConPlex. The motivation for this is that the method of sentence tokenisation in this study seems to represent sentences more accurately than splitting by any occurrence of a period POS tag. In addition, a similar function to safeBreaks was not added, as this is the kind of methodological decision that should be made by individual researchers at the corpus pre-processing stage to match the research questions of the project. The inability to replicate Nelson’s (2024) results precisely is a limitation of the current study. To somewhat overcome this limitation, the *Python* code used in ConPlex is shared on GitHub so that interested researchers can compare it with the *Mathematica* code shared in the supplementary information in Nelson’s (2024) paper.

The results from the *CEFR English Listening Corpus* are shown in Figure 5. The plots represent the distribution of scores throughout the sample. The median is indicated by the solid lines in the boxplots and the mean is indicated by the dotted lines. An incremental increase in complexity scores is evident by examining the boxplots. However, the distributions that are visualised by the violin plots reveal a distinct increase between the B1 and B2 level, and to a lesser extent between the A1 and A2 levels. There are also fewer outliers when compared with the ICNALE data. This could be related to the difference in corpus size, but it might also be related to the more consistent nature of sentence punctuation that exists in transcripts that have been prepared by educational professionals to support learning from listening texts, when compared with written and spoken texts that have been produced by L2 users of the language. Although no previous studies have assessed the complexity of CEFR-aligned listening texts, the results of the current study are somewhat similar to previous research

that investigated the features that discriminated between sentences that were rated from A1 to C2 level. Uchida *et al.* (2024) found that sentences at A and B levels showed lexical and syntactic variation, whereas B- and C-level texts could only be distinguished by lexical aspects. There is greater difference between the A1 and B2 levels in the *CEFR English Listening Corpus*, suggesting that the complexity of constructions is also more variable at the lower CEFR levels.

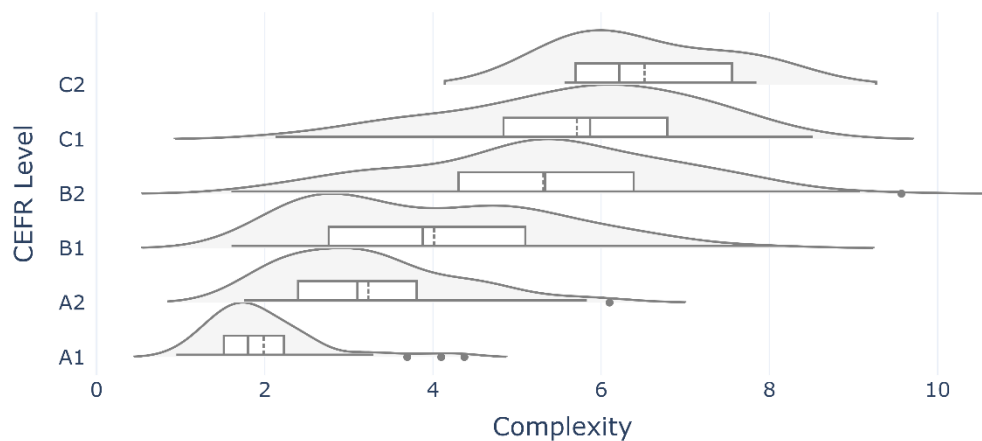


Figure 5: Distribution of complexity scores in the *CEFR English Listening Corpus*

The results from the *MERLIN* corpus are visualised in Figures 6, 7, and 8 for Czech, German, and Italian, respectively. The general trend for German and Italian texts is an increase across CEFR levels. These findings are in line with previous research into L2 German complexity (Weiss and Meurers 2019) that demonstrated accurate text classification from the A2 to B2 level with a selection of 150 complexity features. Classification accuracy was much lower for the A1 and C1/C2 levels, but the general trend aligns with the current study's results, that complexity increases with CEFR level. In L2 Italian, morphological complexity has been shown to be able to distinguish between low- and high-level proficiency learners between the A2 and B2 level (Brezina and Pallotti 2019). However, it was not able to distinguish between the B1 and C2 CEFR levels (Spina 2025). While the current study showed a clear progression, particularly from the A2 to B1+ level, it is not clear whether construction complexity also levels off at the independent to proficient level. Future research could investigate this further with L2 Italian corpora for higher proficiency learners.

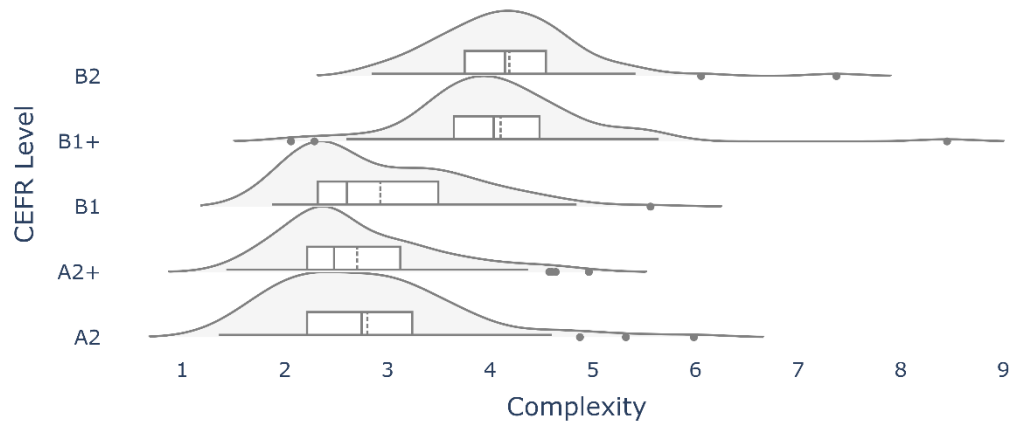


Figure 6: Distribution of complexity scores in the *MERLIN* corpus: Czech

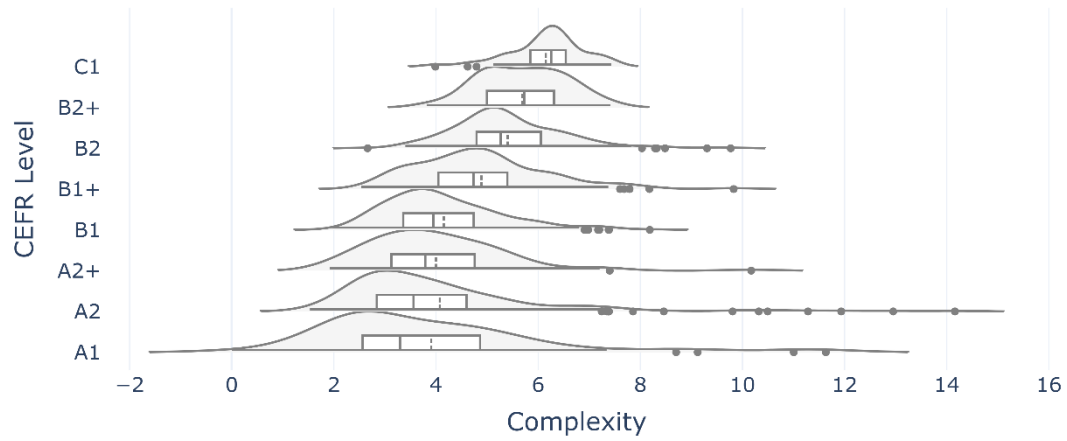


Figure 7: Distribution of complexity scores in the *MERLIN* corpus: German

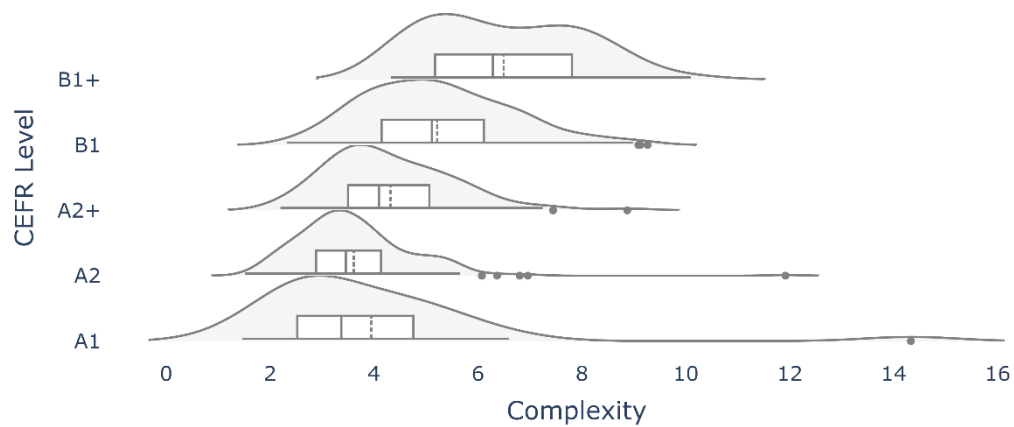


Figure 8: Distribution of complexity scores in the *MERLIN* corpus: Italian

On the other hand, for Czech texts, there seems to be a clear split between complexity scores at the lower levels (A2, A2+, B1) and the higher levels (B1+, B2), whereas the variation within these two groups is more limited. These results somewhat align with previous L2 Czech research (Nogolová *et al.* 2023) that showed a tendency for sentence length, clause length, and number of clauses per sentence to increase from the A1 to B2 level. From the C1 level there was little or no increase in the metrics. Although the researchers pointed out that clause length does not always indicate an increase in syntactic complexity, they argued that clause and sentence length are likely to somewhat correspond with syntactic knowledge. The reason for the difference in CEFR level increase thresholds could be related to the difference in the operationalisation of complexity, or the corpora used in the study. Future research might compare the different complexity measures on the same corpora for further insight into their alignment, or lack thereof.

Overall, the results of the additional datasets analysed in the current study provide further evidence that Nelson's (2024) complexity measure is able to reveal patterns in listening texts in line with the developmental level of the L2 users that they are aimed at. Furthermore, a similar pattern is evident in written texts produced by learners of Czech, German, and Italian.

4.2. Potential uses and limitations

ConPlex has several potential uses, as Nelson's (2024) complexity measure is designed to measure the developmental complexity of language in general, meaning that it is not limited to SLA. It could be used to investigate the complexity of production by first language (L1) and L2 users across developmental levels such as age or CEFR level, as demonstrated by Nelson (2024). In addition, further research could investigate the nature of how construction-based complexity increases across texts that have been produced for L2 reading or listening, as was partially demonstrated in the current study. If a larger corpus was used, a benchmark for each CEFR level could be suggested to measure the complexity of constructions in individual texts. This could provide useful guidelines about the tendencies of text complexity that could be useful for educators and language learners when selecting appropriate texts. The current study also showed that the trend for an increase in construction complexity across CEFR levels extends to languages beyond English. So far, only Czech, German, and Italian have been

considered, but future research could be extended to any of the 70 human languages, at the time of writing, that are supported with pretrained neural models in *Stanza*. It would be worth investigating the constructional complexity of languages in relation to known differences between the languages, such as morphological complexity or word order freedom. In Nelson's (2024) study, the complexity measure was also applied to political speeches. With this in mind, the measure might be of interest to digital humanities researchers if they wish to compare the complexity of constructions used by particular authors, orators, or other language users.

Recently, complexity measures are often integrated into methodologies that aim to assess the readability of texts (Crossley *et al.* 2023), L2 learner writing (Lu 2017), and within the complexity, accuracy, fluency, and lexis framework to measure L2 language performance (Skehan 2009). They can also support the evaluation of interlanguage development over time and provide support in answering other fundamental questions in SLA (Bulté *et al.* 2024). ConPlex could be used to incorporate construction-based complexity into these and other frameworks in the fields of SLA, natural language processing, and beyond.

The main limitation of the tool is its sensitivity to sentence boundaries. How to pre-process texts into sentences is an important methodological consideration that must be made by researchers before using ConPlex. In particular, how spoken texts should be segmented into sentences to represent complexity across utterances is something that should be considered further. Although Nelson's (2024: 23) research showed that the contribution of mode to complexity was small when considering the spoken and written texts in ICNALE, it could be the case that spoken texts have a different complexity threshold to written texts when other corpora are considered. It is possible that the tool could potentially be biased towards measuring complexity in written texts due to its use of the sentence as the unit of measurement. However, these suggestions need to be further investigated in empirical research. Another limitation is the one-dimensional nature of the output of the tool. Depending on the tool's uptake in the research community, further features could be added. For example, complexity scores could be output at the sentence level to allow for more fine-grained analysis and the investigation of complexity across texts. In addition, the tagged sentences and tag pairs for each sentence could be output or visualised, so researchers can gain more insight into the kinds of constructions that are being used across sentences and texts.

5. CONCLUSION

The current study has fulfilled its aim of producing a tool that adequately replicates Nelson's (2024) construction-based complexity measure. Although there were differences in the ICNALE complexity scores between both studies, the way that sentences were operationalised here, using *Stanza*, allows for more accurate calculation of the measure. The creation and release of ConPlex will allow more researchers to experiment with using this complexity measure to answer a range of research questions. The release of the code along with the tool allows for further modifications to make the tool applicable to other languages, as was demonstrated in the current study with Czech, German, and Italian, and texts that require different taggers. It is hoped that the research community will embrace the tool, adding another dimension to the complexity measure debate.

REFERENCES

- Biber, Douglas, Bethany Gray, Tove Larsson and Shelley Staples. 2024. Grammatical analysis is required to describe grammatical (and “syntactic”) complexity: A commentary on “complexity and difficulty in second language acquisition: A theoretical and methodological overview.” *Language Learning*. <https://doi.org/10.1111/lang.12683>
- Boyd, Adriane, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Reykjavik: European Language Resources Association, 1281–1288.
- Brezina, Vaclav and Gabriele Pallotti. 2019. Morphological complexity in written L2 texts. *Second Language Research* 35/1: 99–119.
- Bulté, Bram, Alex Housen and Gabriele Pallotti. 2024. Complexity and difficulty in second language acquisition: A theoretical and methodological overview. *Language Learning*. <https://doi.org/10.1111/lang.12669>
- Crossley, Scott, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi and Agnes Malatinszky. 2023. A large-scaled corpus for assessing text readability. *Behavior Research Methods* 55/2: 491–507.
- Ehret, Katharina, Aleksandrs Berdicevskis, Christian Bentz and Alice Blumenthal-Dramé. 2023. Measuring language complexity: Challenges and opportunities. *Linguistics Vanguard* 9/s1: 1–8.
- Goldberg, Adele E. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7/5: 219–224.
- Hawkins, John A. 2015. *A Comparative Typology of English and German: Unifying the Contrasts*. Abingdon: Routledge.

- Hledíková, Hana and Magda Ševčíková. 2024. Conversion in languages with different morphological structures: A semantic comparison of English and Czech. *Morphology* 34/1: 73–102.
- Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, Adriane Boyd and Henning Peters. 2023. Explosion/spaCy: v3.7.2: Fixes for APIs and requirements. Zenodo. <https://doi.org/10.5281/ZENODO.1212303>
- Ishikawa, Shin'ichiro. 2023. *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English*. Abingdon: Routledge.
- Kettunen, Kimmo. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics* 21/3: 223–245.
- Kyle, Kristopher. 2016. *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-based Indices of Syntactic Sophistication*. Atlanta, GA: Georgia State University dissertation.
- Kyle, Kristopher, Scott A. Crossley and Scott Jarvis. 2021. Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly* 18/2: 154–170.
- Larsen-Freeman, Diane. 1997. Chaos/complexity science and second language acquisition. *Applied Linguistics* 18/2: 141–165.
- Larsen-Freeman, Diane. 2017. Complexity theory: The lessons continue. In Lourdes Ortega and ZhaoHong Han eds. *Complexity Theory and Language Development: In Celebration of Diane Larsen-Freeman*. Amsterdam: John Benjamins, 11–50.
- Larsen-Freeman, Diane and Lynne Cameron. 2008. Research methodology on language development from a complex systems perspective. *The Modern Language Journal* 92/2: 200–213.
- Lu, Xiaofei. 2017. Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing* 34/4: 493–511.
- Lu, Xiaofei. 2024. Towards greater conceptual clarity in complexity and difficulty: A commentary on “complexity and difficulty in second language acquisition: A theoretical and methodological overview.” *Language Learning* <https://doi.org/10.1111/lang.12688>
- MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah: Lawrence Erlbaum Associates.
- McNamara, Danielle S., Arthur C. Graesser, Philip M. McCarthy and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. New York: Cambridge University Press.
- Mizumoto, Atsushi. 2024. Developing and disseminating data analysis tools for open science. In Luke Plonsky ed. *Open Science in Applied Linguistics*. Online: Applied Linguistics Press, 123–131. https://www.appliedlinguisticspress.org/home/catalog/plonsky_2024
- Nelson, Robert. 2024. Using constructions to measure developmental language complexity. *Cognitive Linguistics* 35/4: 481–511.
- Nogolová, Michaela, Radek Čech, Michaela Hanušková and Miroslav Kubát. 2023. The development of sentence and clause lengths in Czech L2 texts. *Korpus - gramatika - axiologie* 28: 22–37.
- Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In Asli Celikyilmaz and Tsung-Hsien Wen eds. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System*

- Demonstrations*. Online: Association for Computational Linguistics, 101–108.
<https://aclanthology.org/2020.acl-demos.14.pdf>
- Shannon, Claude, E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27/3: 379–423.
- Skehan, Peter. 2009. Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics* 30/4: 510–532.
- Spina, Stefania. 2025. Complexity and accuracy of verbal morphology in written L2 Italian: The role of proficiency and contingency. *International Journal of Learner Corpus Research* 11/1: 114–144.
- Uchida, Satoru, Yuki Arase and Tomoyuki Kajiwarara. 2024. Profiling English sentences based on CEFR levels. *ITL - International Journal of Applied Linguistics* 175/1: 103–126.
- Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, and Evgeni Burovski. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods* 17/3: 261–272.
- Weiss, Zarah and Detmar Meurers. 2019. Broad linguistic modeling is beneficial for German L2 proficiency assessment. In Andrea Abel, Aivars Glaznieks, Verena Lyding and Lionel Nicolas eds. *Widening the Scope of Learner Corpus Research: Selected Papers from the Fourth Learner Corpus Research Conference*. Louvain-la-Neuve: Presses Universitaires de Louvain, 419–435.

Corresponding author

Christopher R. Cooper
 Rikkyo University
 Center for Foreign Language Education and Research
 Nishi Ikebukuro 3-34-1, Toshima-ku
 Tokyo
 171–8501
 Japan
 Email: cooper@rikkyo.ac.jp

received: January 2025
 accepted: March 2025

The language of evaluation and stance in crowdfunding project proposals

Alberto A. Vela-Rodrigo
University of Zaragoza / Spain

Abstract – Today, digital crowdfunding platforms allow researchers to increasingly use digital resources to reach and engage diversified audiences, making scientific content accessible to everyone. This paper explores how evaluation in text contributes information relevant to understanding how scientists use language to express their expert opinions of scientific research and their attitudes about the value of their projects. Starting from the compilation and analysis of a 50-science project corpus from Experiment.com, evaluative stance expressions in this work were classified according to Biber's (2004) taxonomy into the following stance categories: verbs, adverbs, adjectives and nouns. Subsequently, genre analysis was applied to identify the discourse functions of these evaluative words in each rhetorical section of the project proposals. Results show that the analysed crowdfunding proposals are rich in stance verbs (52.65%) and, to a lesser extent, stance adjectives (23.52%), serving to express values of effort, improvement and diligence in the proposed projects, as well as judgement regarding experiments and 'Lab Notes' updates, respectively. This can be useful for both theoretical advancement and pedagogical purposes, that is, to apply scientists' findings to digital communication teaching and learning.

Keywords – digital science; evaluation; stance-taking; crowdfunding; genre analysis

1. INTRODUCTION¹

This study explores how the evaluation of texts informs our understanding of writers' stance-taking in crowdfunding scientific projects online. This exploration is deemed important because, as Goźdz-Roszkowski and Hunston (2016: 133) put it, evaluation and stance contribute “to the interactive property of language, to the recognition of how a text is organised, and to the connection between discourse and ideology.” The study of evaluation in texts helps to establish a relationship between linguistic concepts, discourse

¹ The research conducted in this article has been funded by the Spanish Ministry of Economy and the State Research Agency under the project *Digital Genres and Open Science: an analysis of processes of hybridization, innovation and generic interdiscursivity* (PID2019-105655RB-I00) and the project *Genre networks for science communication and dissemination online: Exploring genre interactions and language variation in Web 2.0 (GENNET2.0)* (PID2023-148454NB-I00). I also thank the support of the Government of Aragón to the research group *Comunicación Internacional y Retos Sociales* (CIRES; H16_23R). This article is a contribution to the research conducted in the Institute of Biocomputation and Physics of Complex Systems (BIFI) of the University of Zaragoza and a contribution to the *project Digital Language and Communication Resources for EU Scientists* (DILAN) (2022-1-ES01-KA220-HED-000086749) funded by the European Commission.



and ideas. Analysing evaluative language in crowdfunding proposals online can therefore be a helpful manner to understand how scientists use language in different ways to express their expert opinion of scientific research and their attitudes about the value of their projects. Exploring evaluation and stance also offers insights into how they look at their potential audience and how they engage with them. Evaluation in text is a key dimension when writing or composing a proposal in order to persuade the reader of the validity of the scientific knowledge and how the project can be backed (Millar *et al.* 2023).

The discourse function of evaluation has received increasing interest in the past decades (see Section 2), and it has been approached from different perspectives. Some authors have studied evaluative language use and stance expressions in written academic discourse (e.g., Hunston 2002; Camiciotti and Tognini-Bonelli 2004; Hyland 2005; Biber 2006; de Waard and Maat 2012), and in spoken academic discourse (Mauranen 2003; Swales and Burke 2003). Evaluation and stance have also been investigated in digital genres aimed at promoting public understanding and engagement with science, such as blogs, tweets, and citizen science projects (e.g., Smith 2015; Zou and Hyland 2019; Luzón and Pérez-Llantada 2022). However, and to the best of my knowledge, there are no research studies on the language of evaluation in the digital genre represented by crowdfunding proposals and this is the gap this study seeks to cover. Given that evaluation is key in academic and professional contexts and that its features and resources might differ across genres (Shaw 2003), it is essential for researchers to examine the evaluative dimension of crowdfunding discourse for both theoretical advancement and pedagogical purposes (i.e., to apply their findings to digital communication teaching and learning).

According to previous studies on the phraseology of crowdfunding project proposals from a lexical bundle perspective (Vela-Rodrigo 2023), lexical bundles conveying stance (e.g., *will be able*, *will help us*, *would like to*) are especially frequent in crowdfunding writing, representing more than one quarter of all word bundles of the analysed corpus. Therefore, it is deemed of interest to continue exploring the language of evaluation and stance in a similar corpus of scientific projects for this study. The study of evaluation in crowdfunding writing can help to understand and accept the researchers' perspective when guiding the audience to accept their claims, and thus reflects the social

action that the genre enacts, helping the democratization of science agenda. The research questions set out for the investigation of evaluative language are the following:

RQ1. What language features express evaluation in crowdfunding proposals online?

RQ2. What communicative functions do these features perform in the analysed texts?

RQ3. More broadly, how does evaluation, as a rhetorical strategy, reflect the social action that this genre enhances?

This paper is structured as follows. Section 2 provides a review of the literature on evaluation and stance, with particular emphasis on research within digital genres. The corpus used in this study is introduced in Section 3, along with the methods and tools employed for its analysis. The results are then presented and contextualized through illustrative examples in Section 4, allowing for a discussion of how these findings align with existing literature on stance, as well as the limitations of the present study. Section 5 rounds up the paper with a discussion of the main findings.

2. PREVIOUS RESEARCH

2.1. Evaluation and stance taking

One of the main aspects in the construction of any discourse is how the writer feels intimately about the topic he/she is writing, representing kinds of meaning that might be ‘subjective’ in contrast to the ‘objective’ or ‘factual’ (Goźdz-Roszkowski and Hunston 2016: 133). Defined as “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” (Thompson and Hunston 2000: 5), evaluation in academic and scientific discourse might seem unnecessary, and even contradictory, at first sight. However, as Hunston posited in her doctoral work (1989), and as subsequent English for Specific Purposes (ESP) and English for Academic Purposes (EAP) studies have corroborated, the particular value system shared by scientific writers and their readers, in which emotive or attitudinal language seem to be prohibited, is richer in evaluative meanings than expected by the nature of texts (Mauranen 2002; Jiang and Hyland 2015; Pérez-Llantada 2024a).

The language of evaluation has been studied from many different perspectives or disciplines, ranging from corpus linguistics, systemic-functional linguistics, and cognitive linguistics to sociocultural linguistics, conversation analysis and interactional linguistics. This may explain why many terms developed independently can be covered under the term ‘evaluation’ (Hunston and Thomson 2000), such as stance[taking] (Biber and Finegan 1989), subjectivity (Lyons 1981), sentiment analysis (Turney 2002; Nasukawa and Yi 2003), opinion mining (Dave *et al.* 2003) or appraisal (Martin and White 2005), yet without a no agreed-upon conception of all of them among analysts. Stance taking, the perspective followed in this paper, emphasises the role of the discourse participants’ choice of language to achieve their communicative intentions.

From a methodological approach, some corpus linguists have studied stance to describe specific words or phrases that mark an attitude in a text, such as Biber *et al.* (1999), who divided the category of stance both grammatically and semantically in different sets of words. An important contribution on stance taking in academic discourse is the work of Hyland (2005), who studied the means by which interaction is achieved in academic argument. Hyland’s metadiscourse framework has been particularly influential in the study of evaluation in academic texts (Hyland 1999, 2005). In his analysis of 240 published research articles from eight different disciplines, this author found that writers in the humanities and social sciences took more explicitly involved personal positions when representing themselves and their work than those in the hard sciences. However, all rhetorical choices from both the humanities and scientific fields revealed the writers’ efforts to persuade their audiences of their claims, a finding supported in subsequent publications on metadiscourse and stance taking in texts (e.g., Sancho-Guinda and Hyland 2012; Jiang and Hyland 2015; Hyland and Jiang 2017). On the other hand, stance taking is a frequent activity in language use and has a role in shaping language form (Englebretson 2007), tailoring information and accommodating utterances to the aims of a specific genre. Several studies in corpus linguistics have notably contributed to the description of the lexis and grammar features that convey evaluation and stance in different registers (e.g., Biber *et al.* 1999; Biber 2004, and especially the taxonomy proposed by Biber 2006), with particular attention to adverbials (Conrad and Biber 2000), adjectives and nouns (Hunston and Sinclair 2000 with their ‘local grammar’ for stance taking) and English modals (Thompson and Hunston 2000). Particularly salient is the analysis carried out by Hunston (2007, 2011) to investigate stance quantitatively and

qualitatively, that is, ethnographically. Hunston addresses the question of where in a paragraph stance is articulated (stance location) by exploring concordances for four stance markers in the *Bank of English*. The purpose was to observe multiple uses of the words/phrases *tragedy*, *dramatic*, *to the point of*, *an increasingly accelerated pace* in context using corpus analytical procedures. Hunston proposes that, taken together, explicit (i.e., what is said) and implicit (i.e., what is implicated) stance indicators form the evaluative basis of a given text, given that evaluative meanings are cumulative and occur across phrases in texts. Furthermore, what distinguishes subjective from objective texts is not “the quantity of explicitly evaluative lexical items in each, but the embedding or otherwise of those items in phraseologies, which frequently co-occur with stance” (Hunston 2007: 83). Those phraseologies can be identified intuitively, but since intuition can sometimes be unreliable, a close examination of many examples is required to corroborate such perception on evaluation (Hunston 2007).

2.2. *Stance taking in digital genres*

Analysing evaluation and stance in digital genres of science communication has helped to determine how scientists use language in particular ways when recontextualising specialized information and adapt specialized content to diversified audiences. Evaluation has been studied in academic blogs in the social sciences (Luzón 2012, 2013; Zou and Hyland 2019), *Twitter* (now *X*) discourse (Smith 2015; Luzón 2023; Villares 2023), *Open Laboratory Notebooks* (Luzón and Pérez-Llantada 2022) and online data articles (Pérez-Llantada 2022), among others. In these genres, expressions of epistemic stance (e.g., *it appears that*, *this may be due to*) make arguments and claims tentative. For example, in *Twitter* discourse, tweets are composed using a variety of (linguistic and non-linguistic) expressions of stance and engagement, as shown by Luzón and Albero-Posac (2020) in their analysis of 150 tweets from linguistic conferences. Luzón and Pérez-Llantada (2022) conducted a research case study to analyse the use of language features realising different communicative functions in Spanish research groups on *Twitter*. For the analysis of linguistic forms in a corpus of 600 tweets in different fields of STEM with various communicative functions (e.g., networking, self-promotion, dissemination), they focused mainly on Hyland’s (2005) model for stance and engagement. Their results corroborated previous results by Luzón and Albero-Posac (2020: 46), showing that scientists use evaluative vocabulary (e.g., *amazing talk*, *great talk*) to praise other researchers’ work

and engage in positive public evaluation of their own work. The use of evaluative language has also been reported in relation to processes of knowledge recontextualization in participatory science genres such as citizen science projects. For example, Pérez-Llantada and Luzón (2023: 101) explain that evaluation of content is realised by non-finite clauses (e.g., adverb phrases and non-finite verb phrases encapsulating *to*-infinitive clauses) making overt the researcher's perspective towards the utterances (e.g., *to actually identify this group*; *hopefully the new network will return images*). Other digital genres, such as research blogs, are also rich in evaluative language, which is used to construct credibility online (Rahimpour 2014; Mauranen 2021). At the same time, the language of evaluation in blogs can be traced in digital comments, especially when participants engage in debates and negotiate disagreement as it happens in *Reddit* (Batchelor 2023). Here they can express attitudes (good-bad, positive-negative) through well-known features of digital communication (e.g., 'likes' by clicking on a button; Mauranen 2021: 33).

Some authors analyse evaluation in web-mediated genres such as webpages through the use of rhetorical strategies and linguistic resources that convey stance and persuasion (Askehave and Nielsen 2005), although their approach is equivalent to the one used in this paper, that is, linguistic features realising discourse (evaluative) functions. Other digital genres that are also characterised by the use of evaluative language are crowdfunding projects. In a scientific context of growing interdependence at a global level, the analysis of the language of crowdfunding platforms is interesting. This is a new genre, which stands out for their practical, dynamic and participatory nature, offering researchers the opportunity to share and disseminate their work, while interacting with a non-specialized public. In these projects evaluative markers are used to express an opinion or to make explicit the significance of a project proposal and claim centrality of the research topic, for example, by using adjectival pre-modifiers in complex noun phrases, as explained by Pérez-Llantada (2021a) in her analysis of linguistic features of biomedical projects in *Precipita*, the Spanish platform for crowdfunding science. Similarly, in spoken genres such as TED talks, presenters use stance markers to express judgments and subjectively position themselves (Scotto di Carlo 2014).

3. DATA AND METHODS

3.1. *Corpus description*

For the present study a small-scale corpus of 50 proposals for crowdfunding scientific research was compiled from Experiment.com,² a platform for crowdfunding science across different disciplines. The corpus totalled 140,478 words and considers the information of all sections or tabs in which this website is organised from left to right ('Overview', 'Methods', 'Lab Notes' and 'Discussion'). Every section has distinct functionalities and therefore recall move organization (Vela-Rodrigo 2025). The 'Overview' and 'Methods' sections provide a summary of the methods and procedures to realize the project goals, including a timeline and the pledged amount of money for every project. In the 'Lab Notes' section researchers post updates for their backers in a similar way as blogs also do and interaction with followers is normally reserved for the 'Discussion' tab, a space in which backers and researchers can post their comments and express their gratitude or moral support (for a more detailed description see also Mehlenbacher 2019; Luzón and Pérez-Llantada 2022).³

3.2. *Evaluative model*

A theoretical and analytical model widely spread and used in the study of evaluation is that of stance and engagement developed by Hyland (2005). The model provides a comprehensive and integrated way of examining the means by which interaction is achieved in academic argument. It classifies metadiscourse markers into four categories: 1) hedges, which present information as opinion (e.g., *might*); 2) boosters, which signal involvement with the topic (e.g., *obviously*); 3) attitude markers, which convey the writer's affective stance using stance verbs, adverbs, and adjectives; and 3) self-mentions, which refer to the use of first-person pronouns to present affective information. This model provides a very clear and easy-to-apply codified typology of the resources that writers use to express their positions and has been used to study digital genres (e.g., Zou and Hyland 2019; Luzón and Albero-Posac 2020; Luzón and Pérez-Llantada 2022; Luzón 2023; Villares 2023). Two other models that have also been widely used in the study of evaluation in academic writing are the 'local grammar' by Hunston and Thompson (2000)

² <https://experiment.com>

³ Full details of the corpus of projects can be downloaded at the following link: https://mega.nz/file/CrZkUJiL#1xBgfOyD_o33MkbdI4uJAX_PUyVIGWR_Lxn0NyfBI34

and the corpus-based grammatical investigations of Conrad and Biber (2000) with subsequent applications in studies by Biber (2004, 2006). However, to the best of my knowledge, these models have not yet been applied in the study of digital genres. Biber's (2004, 2006) studies of evaluation and stance in spoken and written academic discourses offer a very versatile taxonomy for the analysis of stance 'content words' (Biber 2004: 123). This model has been chosen for this study since it is very convenient for the analysis of an emerging genre as digital crowdfunding proposal. It offers a very structured and clear framework easy to apply across academic genres, especially considering that crowdfunding proposals borrow certain discursive features with the traditional grant proposal (Mehlenbacher 2017, 2019; Pérez-Llantada 2021b).

Because the present study is exploratory, Biber's (2006: 112) taxonomy was simplified and adapted for the analysis, including groups of similar categories under the same general term (e.g., modal verbs have been combined with controlling verbs in *to/that* clauses). The resulting taxonomy classifies stance words according to their grammatical domain and semantic function considering markers of stance in *that/to*-clauses together, according to every word category, that is, stance adverbs, stance verbs (including (semi-)modals), stance adjectives and stance nouns. This adapted taxonomy (Table 1) is more appropriate to apply in a small corpus such as the one used in the present study, especially when it comes to offering comparative data across rhetorical sections.

Stance Category	Subcategories	Examples
Stance Verbs	- Modal / Semimodal verbs	<i>can, could, may, will</i>
	- Attitude/Intention/Desire	<i>expect, intend</i>
	- Non -factive/Communication speech verbs	<i>address, relate, inform</i>
	- Factive verbs (certainty)	<i>know, ensure</i>
	- Effort/Facilitation	<i>allow, help, support</i>
	- Likelihood/Cognition	<i>estimate, consider</i>
Stance Adverbs	- Attitude (evaluation/expectation)	<i>amazingly, importantly</i>
	- Certainty	<i>certainly, in fact</i>
	- Communication speech	<i>additionally, finally</i>
	- Likelihood	<i>perhaps, probably</i>
Stance Nouns	- Epistemic/Attitudinal	<i>interest, love, success</i>
	- Certainty	<i>evidence, expertise</i>
	- Communication speech (including prepositional/noun phrases)	<i>in addition, a bit</i>
	- Likelihood	<i>hypothesis, condition</i>
Stance Adjectives	- Epistemic/Attitudinal	<i>good, bad</i>
	- Certainty	<i>certain, sure</i>
	- Communication	<i>explicit, informative</i>
	- Likelihood	<i>probable, possible</i>
	- Evaluative	<i>important, beautiful, interesting</i>

Table 1: Taxonomy of stance grammar particles adapted from Biber (2006)

The quantitative results of applying this taxonomy are based on raw counts, but, since the rhetorical sections contained a different number of words, the raw frequencies were normalised per 1,000 words to carry out the comparison across sections.

3.3. Identification of evaluative language

Regarding the analysis of stance features in this study, evaluative stance expressions were first identified through the extraction of content words and then classified manually into the following categories: verbs, adverbs, adjectives and nouns. Afterwards, they were filtered by selecting only those conveying evaluative meanings. The details of all the steps followed are detailed below.

The first step involved the automatic extraction of 15,781 content words (66,818 tokens) from the corpus. According to Biber (2004, 2006), content words are those that can convey evaluative meanings, therefore limiting the scope of the search was important. To identify the different types of content words *Lancsbox* 3.0.0 (Brezina and Platt 2023) was used, which allows automatic tagging of the corpus texts for different grammar categories previously converted to .txt format. This software offers an advanced search tool called *Words Tool* that filters the different lemmas or semantic domains of the words according to their frequency and dispersion. This tool was used to generate lists of word types occurring in each rhetorical section of the proposals.⁴

Table 2 shows the total counts (types) and tokens of each content word and their distribution in each section. This preliminary step enabled the extraction of content words that could potentially contain evaluative meanings, the analysed text representing 81.71% of all words in the corpus. The remaining word types correspond to other categories such as pronouns (e.g., *I, this, mine*), prepositions (e.g., *under, against, of*), determinants (e.g., *the, a*) conjunctions (e.g., *and, or*), and interjections (e.g., *oh!, wow*), were not considered for this study since they are not content words.

⁴ Lists downloadable at <https://mega.nz/folder/DzpAlaAD#DwGhjIzJSoooHWIKYckNYg>

Category/ Section		Overview		Lab Notes		Discussion	
		raw	norm	raw	norm	raw	norm
verbs	types	882	6.27	1,124	8	391	2.78
	tokens	6,264	44.59	7,953	56.61	1,857	13.21
adverbs	types	309	2.19	410	2.91	156	1.11
	tokens	1,669	8.32	3,170	22.56	780	5.55
adjectives	types	1,006	7.16	1,163	8.27	331	2.35
	tokens	4,995	35.55	5,391	38.37	1,109	7.89
nouns	types	3,753	26.71	4,794	34.12	1,462	10.4
	tokens	14,836	105.61	16,174	115.13	2,620	18.65

Table 2: Raw and normalised counts (per 1,000 words) of content word types and tokens across rhetorical sections

The resulting lists were subsequently checked and filtered manually to identify only those potentially evaluative words that conveyed evaluative meanings. For the identification of these evaluative words in context, I relied on Biber's (2006) taxonomy and examples of previous classifications (Biber 2004: 133–135). Biber's taxonomies offer a very intuitive and explanatory interpretive framework along the same lines of Hunston and Thompson's (2000) local grammar, allowing retrieved data to be compared under premises, a method which has already proven to be useful for examining evaluation. For example, among the potentially evaluative word contents in this first checking, it is possible to find adjectives (e.g., *different*, *good*, *important*); nouns (e.g., *love*, *opportunity*, *hope*); adverbs (e.g., *hopefully*, *sadly*); and verbs (e.g., *help*, *like*, *feel*).

This manual process was assisted by the KWIC tool in *Lancsbox* 3.0.0, which helped to retrieve concordance lines and show the words that might convey evaluative meaning in its context. In other words, the KWIC tool allowed context-sensitive analysis. Grammars also informed the selection of evaluative markers. For example, according to Biber (2006), adjectives which act as a controlled word in *that*-clauses had an evaluative meaning; therefore the word (e.g., *different* in example (1) below) should be checked in context in order to see whether it is accompanied by prepositions such as *from* or any other phrase category, as in (1).

- (1) micro-residues recovered within the intentional fires are **different** from the micro-residues recovered outside the fires.

Since in this sentence *different* does not act as an adjective controlling a *that*-clause, this token can be removed from the list of adjectives. Similarly, verbs can be controlled words

in *that*-clauses and *to*-clauses, which means that it is necessary to check whether every verb in the preliminary list (e.g., *hope*) takes part in such clauses, as in (2).

- (2) By examining the composition of shells, I **hope** to discover how variations in stable isotopic.

In this case, *hope* is part of a *to*-clause, therefore the verb has an evaluative meaning and can remain in the list. Table 3 shows the percentage of non-evaluative versus evaluative content words.

		Section		
		Overview	Lab notes	Discussion
Verbs	Non. eval.	3,249 (51.87%)	4,856 (60.16%)	1,070 (57.62%)
	Eval.	3,015 (48.13%)	3,097 (39.84%)	787 (42.38%)
Adverbs	Non. eval.	1,198 (71.78%)	2,634 (83.1%)	571 (73.21%)
	Eval.	471 (28.22%)	536 (16.9%)	209 (26.79%)
Adjectives	Non. eval.	3,976 (79.59%)	4,463 (82.78%)	559 (50.41%)
	Eval.	1,019 (20.41%)	928 (17.22%)	550 (49.59%)
Nouns	Non. eval.	13,937 (93.93%)	15,531 (96.03%)	2,287 (87.3%)
	Eval.	902 (6.07%)	643 (3.97%)	333 (12.7%)
Totals	Non. eval.	12,516 (69.83%)	27,484 (4.08%)	4,487 (70.49%)
	Eval.	5,407 (30.17%)	5,204 (15.92%)	1,879 (29.51%)

Table 3: Distribution of non-evaluative and evaluative content words across rhetorical sections

The resulting lists of evaluative words amounted to 12,490 tokens, distributed as shown in Figure 1.

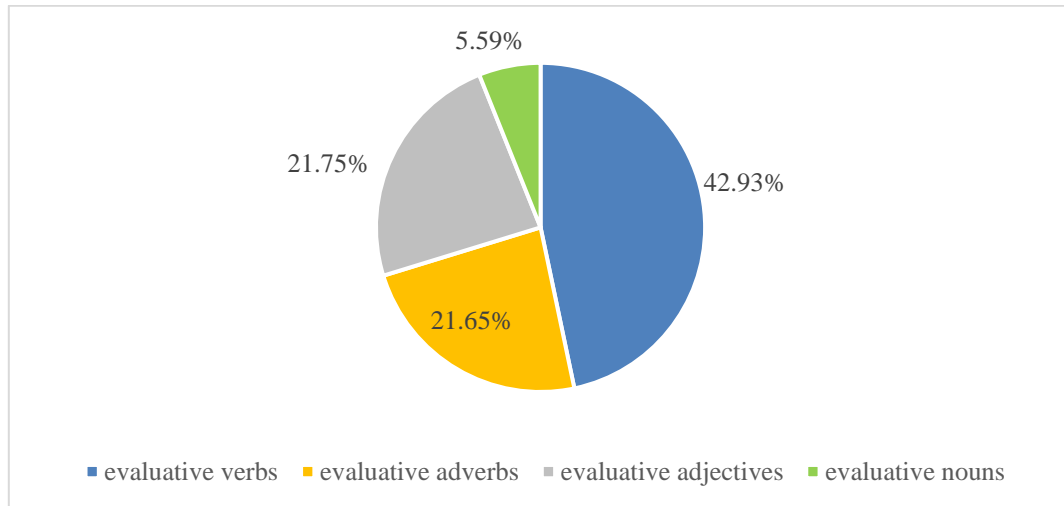


Figure 1: Distribution of evaluative words according to their grammar category

These evaluative words comprise more than 20% of all words analysed. Thus, considering the total number of 140,478 words the corpus contains, evaluative words represent 8.76% of all text in the corpus.

Regarding every rhetorical section (Figure 2), 5,407 words have an evaluative meaning in the ‘Overview’ section (30.17% of all words in that section), 5,592 words (15.92%) in ‘Lab Notes’ and 1,879 words (29.51%) in the ‘Discussion’ section.

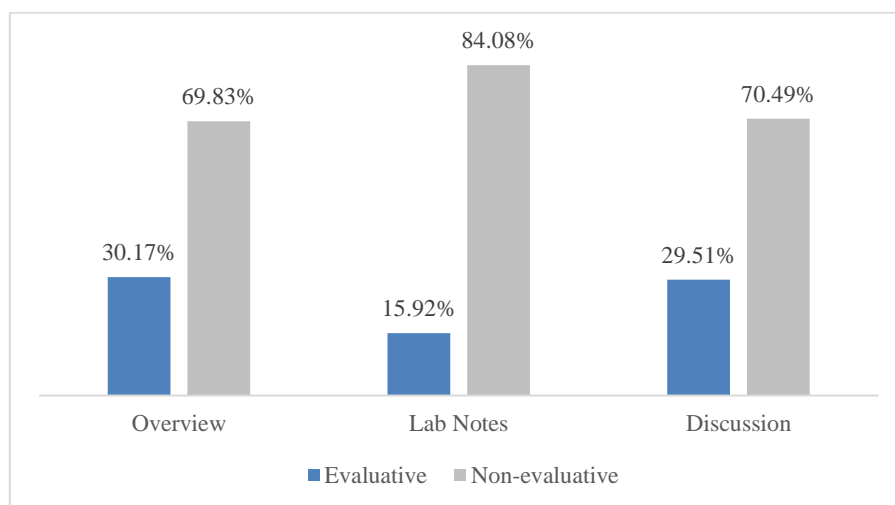


Figure 2: Proportion of evaluative and non-evaluative words in every rhetorical section

These evaluative words were further classified in charts according to Biber’s (2006: 112) taxonomy for spoken and written registers in academic discourse, which is a result of previous works by the same author (e.g., Biber *et al.* 1999: 353–388 and Biber 2004). This analytical framework was chosen in order to understand the functions of evaluative words in the crowdfunding proposals. Biber’s framework classifies stance words according to their grammatical domain and semantic function and distinguishes the following major structural types of stance grammatical markers: i) modal / semi-modal verbs, ii) stance adverbs, and iii) stance adjectives, verbs or nouns acting as controlling elements in complement clauses (*that/to*-clauses). According to Biber (1999: 967) “grammatically marked stance is the most overt manner to express stance, over value-laden word choices or paralinguistic devices”. To interpret these results, the framework of genre theory was used (Swales 2004), together with previous studies on digital genres and academic and professional discourses for genre analysis (Askehave and Nielsen 2005; Mehlenbacher 2017, 2019; Luzón and Pérez-Llantada 2022).

4. RESULTS AND ANALYSIS

4.1. Overall findings

12,490 evaluative words were classified according to their grammar category (Biber 2004, 2006).⁵ Table 4 shows the proportion (%) of each grammatical subcategory in the total of evaluative words found in the corpus (verbs, adverbs, adjectives, nouns).

Stance Category	Subcategory	%	Examples
Stance Verbs	Modals and Semimodals	33.94%	<i>shall, may, be going to</i>
	Attitudinal/Intentional	19.86%	<i>expect, intend, aim</i>
	Effort	19.06%	<i>accomplish, facilitate</i>
	Factive	11.55%	<i>demonstrate, ensure</i>
	Non-factive	8.00%	<i>assure, assume</i>
	Cognition	7.59%	<i>think, seem</i>
Stance Adverbs	Attitude / Personal	39.00%	<i>extremely, simply</i>
	Affect		
	Non-factive (Speech)	29.85%	<i>enough, especially</i>
	Factive	19.07%	<i>absolutely, eventually</i>
	Likelihood	12.08%	<i>usually, apparently</i>
Stance Adjectives	Evaluative	65.03%	<i>amazing, dangerous</i>
	Attitudinal/Intentional	14.36%	<i>amazed, aware</i>
	Ability	10.00%	<i>able, capable</i>
	Likelihood	5.11%	<i>current, likely</i>
	Factive	5.50%	<i>true, worthy</i>
Stance Nouns	Attitudinal	43.81%	<i>abundance, attitude</i>
	Factive	22.66%	<i>conclusion, significance</i>
	Likelihood	20.23%	<i>chance, hypothesis</i>
	Non-factive / Speech	13.20%	<i>in addition, a bit</i>

Table 4: Proportion (%) of each grammatical subcategory in the total of evaluative words found in the corpus

If we turn to the general data of stance words in the corpus, Figure 3 summarises the salience of stance features according to their grammatical domain, which indicates that the use of stance verbs (6,899 words; 55.23%) and, to a lesser extent, stance adjectives (2,497 words; 19.99%) are particularly important in these proposals. The fact that more than half of the stance words in the texts are verbs is significant, although it must be highlighted that many of them correspond to modal verbs (2,346 words; 18.78%), expressing possibility/permission/ability (*can, could, may, might*), logical necessity/obligation (*must/should*), and prediction/volition (*will/would/shall/be going to*).

⁵ Lists of stance words downloadable at https://mega.nz/file/nrY3BIBS#HxAcdf5cipyqUtwbbYZKfaj2_0SkUYx0ffE_FGos9C_8

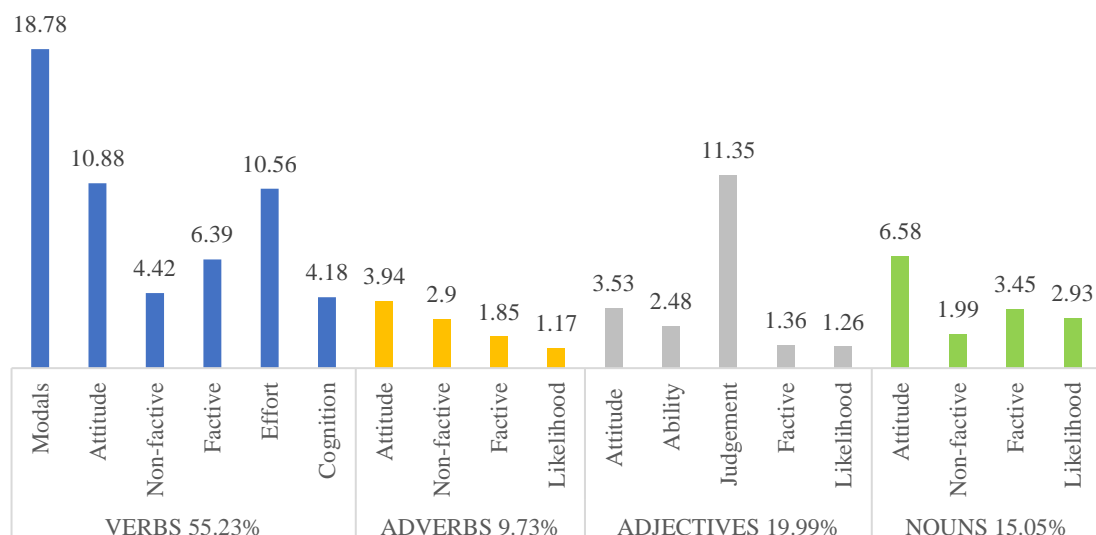


Figure 3: Distribution (%) of evaluative words expressing stance in the corpus (all sections)

Modal and semi-modal verbs are very common in conversation (Biber 2004) and while modal verbs (e.g., *shall*, *might*) have been generally on the decline over the past decades, selected modals (e.g., *can* and *will*) and selected semi-modals (e.g., *be going to*) are used with increasing frequencies in English (Biber *et al.* 1999: 221–262), as shown in examples (3) and (4).

(3) Hi Cindy, Thanks for asking! I **will** isolate the foraminifera from the surrounding sediment by disaggregation and sieving. [DOI: 10.18258/6079]

(4) Want more details about the finds we've already made at the AAS and what we hope to accomplish as research moves forward? You're in luck, because we're **going to** be participating in a Reddit Ask Me Anything event tomorrow at 1:30 PM [DOI: 10.18258/6865]

Thus, as in other genres based on written communication (e.g., text books, university catalogues and brochures; Biber 2006) verbs to mark attitudes/desire (e.g., *appreciate*, *become*, *concern*) (1,359 words; 10.88%) and to express effort (e.g., *accomplish*, *manage*, *allow*) (1,319 words; 10.56%) constitute a significant portion of all stance verbs in the corpus, controlling both *that*-clauses and *to*-clauses (e.g., *we expect that*; *we intend to*; *we encourage to*). These stance verbs serve to express values of effort, improvement and diligence in the proposed projects, while scientists position themselves with respect to their own commitments, humanizing the fundraising and research process, expressing their opinions, feelings and doubts, as illustrated in examples (5) to (7).

- (5) As a soil scientist I **appreciate** knowing about research projects too. [DOI: [10.18258/11434](https://doi.org/10.18258/11434)]
- (6) We **encourage** you to reproduce and adapt these designs to address your own unique environmental monitoring needs. [DOI: [10.18258/7455](https://doi.org/10.18258/7455)]
- (7) Hopefully this project **will** shed more light on whether or not global warming actually exists. If I were a betting man though, I'd **expect** that this whole thing is a liberal conspiracy. [DOI: [10.18258/7455](https://doi.org/10.18258/7455)]

On the other hand, the presence of stance adjectives (almost 20%), which are especially common as controlling words of *to*-clauses (e.g., *it is possible to decide*; *it is difficult to establish*) and are typical of written registers (Biber 2006), serve primarily to express judgements (1,418 words, i.e., 11.35%; e.g., *inappropriate*, *lovely*, *delicious*), as shown in examples (8) and (9). These judgements are based on the personal values and beliefs of the scientists writing each proposal and, in many cases, may be values shared by the entire scientific community. Judgement adjectives comprise positive (e.g., *good*, *valuable*) and negative (e.g., *bad*, *wrong*, *poor*) adjectival evaluation (Thompson 2014).

- (8) Jim, **lovely** to have you on board. If you are interested in coming down and helping with field work, or just visiting, we'd love to have you! [DOI: [10.18258/6913](https://doi.org/10.18258/6913)]
- (9) Every dig they work **hard** to get us there with every shovelful of dirt, every trowel-turn of sediment, every single day spent uncovering an ancient Cretaceous coast. [DOI: [10.18258/6865](https://doi.org/10.18258/6865)]

As shown in Figure 3, attitude stance nouns are also relatively common, amounting 6.58% (1,878 words), and occurring especially in *that*-clauses (e.g., *the expectation that creates*; *the support that we received*). *That*-clauses controlled by nouns are restricted primarily to the academic registers and in them the nouns serve to identify the status of the information presented in the clauses (Biber 2006), as in example (10).

- (10) The **idea** that humans were interacting with the Warrah's ancestor, lends itself well to the idea that perhaps the Warrah is a remnant semi-domesticated form of its extinct ancestor. [DOI: [10.18258/3682](https://doi.org/10.18258/3682)]

Therefore, it can be deduced that the academic register plays an important and characteristic role in this type of projects, since the information they present is similar to that also presented in the antecedent of this digital genre, that is, the grant proposal (Mehlenbacher 2019). From the findings, attitudinal adverbs (e.g., *correctly*, *perfectly*) (493 words; 3.94%) and non-factive adverbs (e.g., *generally*, *ideally*) (363 words; 2.9%) also seem to be important in these projects. The former conveys an assessment of

expectations when carrying out the different steps of a project or the convenience and future application of the research, as in (11) about the faunal diversity in the Mesozoic formations of Northwestern Colorado. The latter comments on the manner of conveying the scientific data of the projects or the perspective that the information is given from (12).

(11) We would greatly appreciate your support but do understand if this research does not **perfectly** fit what you are looking for. [DOI: 10.18258/12864]

(12) **Generally**, soil fertility decreases, the amount of organic matter increases, and soil texture becomes finer with forest succession. [DOI: 10.18258/6913]

4.2. *Comparison of stance functions across sections*

Corpus findings also show distinctive stance words in each rhetorical section ('Overview', 'Lab Notes', 'Discussion').⁶ To understand these results, it is necessary to carry out an analysis by sections that allows us to understand the representativeness of the stance words in their context and the discursive strategies they imply.

4.2.1. 'Overview' section

Figure 4 shows a similar tendency to that already observed in the corpus analysis of the general data (Figure1). In the 'Overview' section stance is carried out mainly through evaluative verbs, that is, modal and semi-modal verbs (1,068 words; 19.75%), followed by verbs of effort (e.g., *require*, *challenge*, *help*; 645 words; 11.92%) and attitudinal/purpose verbs (e.g., *aim*, *hope*, *want*: 418 words; 7.73%).

⁶ Lists of stance words in each rhetorical section downloadable at:
https://mega.nz/file/vvY1HJ4b#pY4KNgFouBc_plI3J32DkF5v7EpkYlrx5FJORgGSB29s

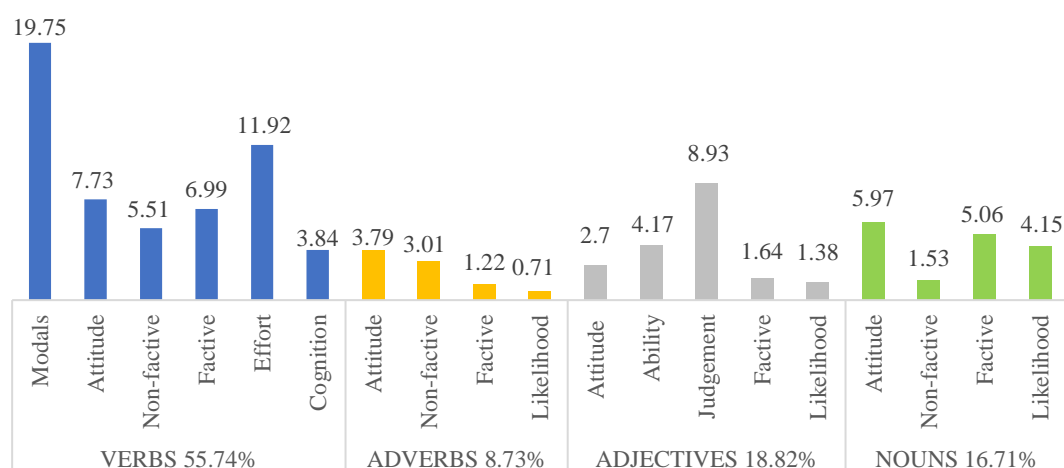


Figure 4: Distribution (%) of evaluative words expressing stance in the 'Overview' section

Their high recurrence indicates the importance of showing the desire, expectancies and/or determination to carry out the aims of the projects, overcoming difficulties during the collection of samples and data or due to lack of material means. This is obvious in subsections such as 'What are the goals of the project?' or 'Additional information' in which researchers can deepen in the significance and context of the project, as in (13), about the use of caves by animals in Southwest Ohio since the Pleistocene.

- (13) Additionally, it is our **hope** that radiocarbon work for one site can help to illustrate the significance of the fossils here, which in turn can be used as a spring board for grant applications further down the road. [DOI: [10.18258/11485](https://doi.org/10.18258/11485)]

Factive verbs (e.g., *know*, *ensure*, *demonstrate*) (378 words; 6.99%) and non-factive verbs (e.g., *predict*, *relate*, *answer*; 298 words; 5.51%) are also used with rather similar frequencies. Normally, scientists rely more often on non-factive verbs to report information and neutrally inform readers of their own position, whereas factive verbs are more commonly used to support their own opinions (Hyland 2002). Therefore, it can be asserted that the use of both evaluative verb categories in the 'Overview' section implies a balance between the presentation of scientific data in an objective manner and the writer's own position towards them, for example, through the acceptance of the results or the potential conclusions of the projects to be crowdfunded, with verbs such as *involve*, *show* or *demonstrate* (see (14)).

- (14) This project is important because its results will **show** that the simple, effective composting system could be replicated in other locations. [DOI: [10.18258/11485](https://doi.org/10.18258/11485)]

From Figure 4 it is also evident the use of judgement adjectives in this section (e.g., *dramatic, relevant, great*; 483 words; 8.93%), similarly to the general tendency observed in the rest of sections of these projects. Expressing personal values and ideas is particularly common in areas such as ‘Meet the Team’, in which researchers introduce themselves sometimes even writing about their hobbies or childhood, or in ‘What are the goals of this project?’ where the researcher (project launcher) can express his/her opinions about the importance of reaching the aims of their research (see (15)).

- (15) This project has **great** potential to restore the native flora through a promising method of strawberry guava removal. [DOI: 10.18258/8423]

Also, rhetorical subsections such as ‘Endorsed by’, in which researchers receive recommendations from colleagues, and ‘Additional information’ are abundant in stance adjectives for judgment, normally controlling that/to clauses. This is exemplified in the following extract from the ‘Additional information’ section of the ‘Overview’ text of a project about soil contamination and the presence of women in the STEM workforce of Nigeria (see (16)).

- (16) [...] This is a thing of notable concern because it is **difficult** in the present-day society to address issues of national development without recourse to gender factor [DOI: 10.18258/20466]

On the other hand, the use of attitudinal nouns (e.g., *effort, support*; 323 words; 5.97%) and factive nouns (e.g., *evidence, effect*; 274 words; 5.06%) is not unexpected in this section, since discourse here presents features of the scientific discourse (i.e., objective data using an academic style in rhetorical subsections such as ‘About this project’) as in (17), combined with the expectations and justifications of the writers towards their project aims, future results or accommodation of their research in the international scientific scene (18).

- (17) In addition, there is growing **evidence** that species diversity and composition are linked to ecosystem function in managed and natural systems, although the mechanisms behind these relationships are debated. [DOI: 10.18258/6740]

- (18) Baltimore is beginning a wide-scale **effort** to climate-proof the city by planting more trees, installing white roofs, and other green infrastructure. Understanding urban temperature and micro-climates can help city planners [...] [DOI: 10.18258/7455]

4.2.2. ‘Lab Notes’ section

‘Lab Notes’ section presents similar results to those of the ‘Overview’ section/tab, with stance verbs still conveying most of stance communicative functions in the corpus (Figure 5). This is still true with the greater use of modal verbs (961 words; 18.56%) over the rest of word categories.

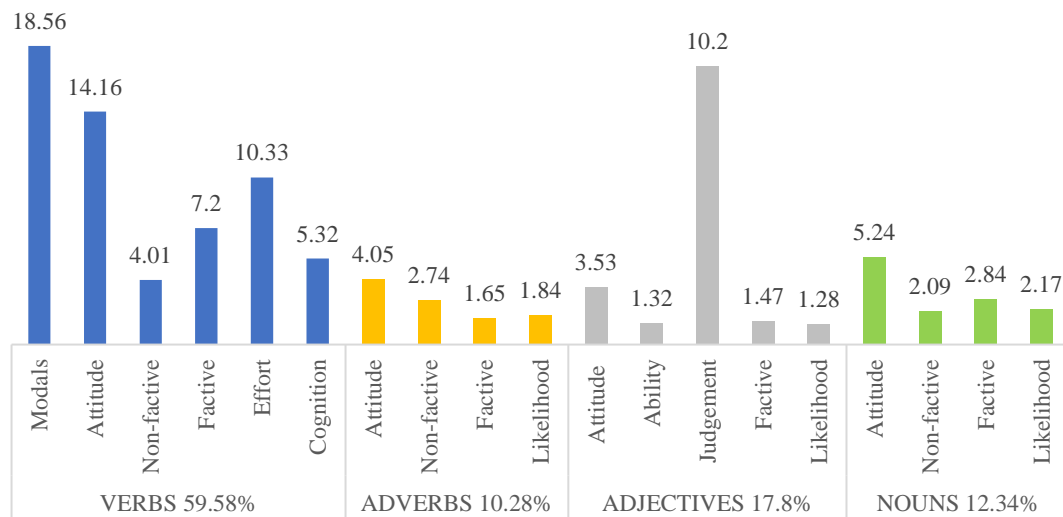


Figure 5: Distribution (%) of evaluative words expressing stance in the ‘Lab Notes’ section

In this section, scientists prefer attitudinal/purpose verbs (e.g., *hope*, *find*, *appreciate*; 737 words; 14.16%) to stance verbs for conveying effort (e.g., *manage*, *encourage*; 538 words; 10.33%). Whereas in the ‘Overview’ section expressing motivation, hard work and dedication are the most important aspects to convince the audience about the determination to reach the aims of the project, in the ‘Lab Notes’ section, intention and desire seem to prevail. This could be explained by the fact that this section consists of posts published as blogs would do, in which updates on specific goals in the development of the project are reported. Most of these posts narrate methodological procedures or small objectives and the purpose behind them, as in example (19) about a project dealing with the destruction of the Middle Bronze Age Civilization in North of the Dead Sea by fire.

- (19) The four proposed coring sites were selected for the reasons stated in my answer my brother's question, but they were also chosen because they are away from (but still near) major wadis. I was trying to **find** locations that will have relatively undisturbed accumulations of alluvial strata. [DOI: [10.18258/6832](https://doi.org/10.18258/6832)]

Adjectives to express judgement (e.g., *valuable*, *wonderful*, *unique*) continue to be particularly relevant in this section, accounting for 10.2% (531 words). This stance category is used to examine the quality of results in the process of implementing the project and to consider and debate the benefits, as in example (20) about the quality of indoor air in Northeast Denver.

- (20) What is exciting to me, is the mutually beneficial nature of CBPR. In addition to results potentially interesting to the academic community, the research can produce outcomes **valuable** to the participating community. [DOI: [10.18258/5329](https://doi.org/10.18258/5329)]

The use of attitudinal nouns in ‘Lab Notes’ (e.g., *interest*, *gratitude*, *effort*) accounts only for 5.24% (273 words) of all stance words. Attitudinal expressions are much more common in speech than in writing (Biber 2006; Pérez-Llantada 2021a; Vela-Rodrigo 2023), which would confirm the presence of conversational elements typical of spoken discourse and informal interaction in this section. Example (21) about the fauna of the floating islands in the Sargasso Sea illustrates this fact.

- (21) Yet here, we have a frogfish living in the middle of the ocean. I hope that you will support my **effort** to understand the differences between those animals living in Sargassum in the Gulf, Sargasso Sea, and Caribbean. Please check in for frequent updates from the lab and the field! [DOI: [10.18258/4746](https://doi.org/10.18258/4746)]

4.2.3. ‘Discussion’ section

The presence of evaluative words expressing stance in the ‘Discussion’ section does not differ significantly from their incidence in the ‘Overview’ and ‘Lab Notes’ sections (Figure 6). However, in this case, adjectives to express judgement (e.g., *amazing*, *impressive*, *unfair*) are particularly prominent (404 words; 21.5%) among the remaining stance particles, being the most frequently used category of stance words in the whole section.

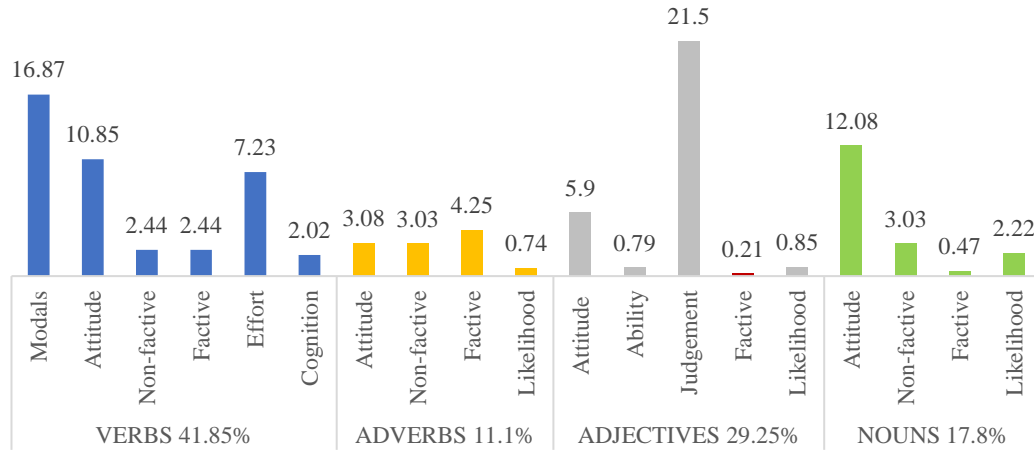


Figure 6: Distribution (%) of evaluative words expressing stance in the 'Discussion' section

This result is not unexpected considering the communicative function of this section, namely the exchange of ideas between writers and backers (in both directions: backer-writer/writer-backer). In here, people express their feelings and opinions openly (example 22), something that is also reflected in the prominent use of attitudinal verbs (204 words; 10.85%) (e.g., *love, find, miss, admire*, as in (23)) and modal verbs (317 words; 16.87%) to express ability and willingness (e.g., *can, might*, as in (24)).

(22) Thank you Jim! Yes, I also think it's **unfair** that future generations may not get to experience nature the way they could have [DOI: 10.18258/6664]

(23) Dear Peter than you ofr (sic) your supportive comments, i am glad you found the interview worthwhile. it is such an exciting project that I **love** telling people about it (sic). [DOI: 10.18258/12850]

(24) Keep up the good work! I **can't** wait to see if you **can** do it. [DOI: 10.18258/8220]

As in any other discussion fora, it is common to find formulas of courtesy or politeness (e.g., *dear, thanks, thank you, with love*). This colloquial discourse features foster familiarity and proximity with backers, making them participant of the research project while scientists receive moral support from backers. On the other hand, verbs to express effort (e.g., *support, achieve, manage*) also play a considerable role in this tab, in which backers encourage scientists to reach their goals and complete their research in time, while at the same time letting them know about the importance of their projects for society. In this section, the linguistic resources perform their communicative function of persuading the audience to support the project in a more subtle manner, since it is through gratitude and courtesy that this task is performed. For that, attitudinal nouns (e.g., *luck, success, passion*) also play an important role in this section, amounting to approximately

12% (227 words) of all stance particles, with backers complimenting scientists or expressing their best wishes, as in example (25).

- (25) This is very exciting to help support real scientific work. Better still is I'll be helping one of my hero's to go into the field to witness it first hand! Best of **luck**, and I hope you guys uncover something grand. [DOI: [10.18258/12850](https://doi.org/10.18258/12850)]

The language is therefore characterised by colloquial features since the virtual space for blogging fosters conversation (with the expected markers of orality) in the writer-backer and backer-writer directions.

5. DISCUSSION AND CONCLUSION

The purpose of this study was to explore how evaluation in text contributes information relevant to understanding how writers communicate their attitudes and opinions and take stances when crowdfunding their scientific project proposals in online platforms. For this purpose, the stance words used in the texts (verbs, adverbs, nouns, adjectives) were identified to understand how they contribute to the organization of a text. The study also intended to widen our knowledge on the different ways used by scientists to express their expert opinion of scientific research and their attitudes about the value of their projects. To this purpose, the article revolved around three main research questions, whose answers have been addressed in light of the data obtained.

In response to RQ1 (What language features express evaluation in crowdfunding proposals?), the findings indicate that more than a half of all evaluative words are stance verbs (55.23%) and, to a lesser extent, stance adjectives (almost 20%). The fact that stance adjectives occurred in crowdfunding proposals suggests similarities between crowdfunding proposals and genres of oral discourse. For example, some adjectives in the texts analysed (e.g., *pretty*, *wonderful*, *crazy*, *awesome*) could be considered a more colloquial option to the formal adjectives typically found in academic writing. Another similarity with spoken discourse was the high use of modal and semi-modal verbs, very common in conversation (Biber 2004). Interestingly, whereas some adjectives occurring in my corpus as controlling words of *to*-clauses were very typical of written and more formal registers (Biber 2006), many of the most frequent adjectives in the judgement and attitude categories were positive adjectives (e.g., *beautiful*, *amazing*, *cool*, *positive*) not so expected in the impersonal style of academic writing (Thompson 2014). Therefore, the crowdfunding proposal integrates different types of discourses according to the

affordances and constraints of the medium and the types of interaction they support (e.g., readers' comments and responses, 'Lab Notes' updates). As shown in Section 4.1, attitudinal expressions were very common in the corpus (almost 25% of all stance words), a typical feature of speech compared to writing contexts (Biber 2006), which indicates that the genre investigated appears to be characterised a hybrid discourse style with both elements of written and oral discourse referring a certain colloquialization of the academic discourse. This fact would reflect the social action that the genre enacts, helping the democratization of science agenda that is making science more participatory to a broader audience (Follet and Strezov 2015). This may also explain that the prevalence of *to*-clauses over *that*-clauses in the corpus also has a clear functionality: explaining or making explicit the purpose of the project or the purpose of the activity carried out by the researcher.

Concerning how stance is conveyed in the different sections ('Overview', 'Lab Notes', 'Discussion'), the analysis displayed similarities among them. They all relied on modal verbs primarily, the use of *will* being especially relevant to explain with certainty how their expected scientific contribution will address the project goals. In addition, the modal *will* also served to convey "immediacy of action and intentionally, or willingness to move the project forward" (Luzón and Pérez-Llantada 2022: 125). The presence of these features is not surprising, as one of the main communicative purposes of the discourse of crowdfunding proposals is to build credibility and ultimately to persuade their audiences to support their research with donations. However, whereas the less formal discourse in the 'Lab Notes' and the 'Discussion' sections involved a profuse use of attitudinal verbs, the 'Overview' section was richer in verbs of effort. The researchers created a persuasive appeal for their potential backers transmitting courage, tenacity and determination to carry out the project with endeavour. Also in the 'Overview' section, stance adjectives of judgement were used to claim significance of the project, being especially relevant in the more colloquial 'Discussion' and 'Lab Notes' sections. The findings showed that these adjectives also served to transmit the personal values of the writers, helping to create proximity with the backers, as also happens in tweets written by scholars and researchers that aim to engage readers (Luzón 2023). It is also worth highlighting the use of attitudinal stance nouns (controlling *to*-clauses) in all the three sections, especially with a positive meaning. The expression of evaluative judgment through nouns helps to persuade readers of the writers' right to speak with authority and

to establish their reputations (Jiang and Hyland 2015), especially important in the ‘Overview’/‘Methods’, whereas the prominent use of attitudinal nouns in the ‘Discussion’ sections seemed to respond to the brevity of the sentences here based on greetings, good wishes and brief comments. This is also constrained by the limited space of the medium in this section, similarly to other digital genres such as *Twitter* (Villares 2023).

Regarding RQ2 (What communicative functions do these evaluative words perform in the texts analysed?), the corpus data showed that in these proposals stance adjectives to express judgements alongside those to express ability were used rhetorically to construct the identity of the project launchers, especially in the ‘Overview’ of the project. In this section, researchers need to answer questions such as ‘What is the context of this research?’, presenting themselves as experts in ‘Meet the Team’, or creating an emotional bond with the audience through a biographical story telling. Since emotive stance particles, especially adjectives, are rare in academic writing, the more academic nature of the ‘Overview’ section, in which scientific data are presented as they would be expected in canonical scientific papers, tend to be “institutionalized” (Martin 2000: 155) semantic choices of emotional values as judgement values in online proposals, as Scotto di Carlo (2014) also reports for the case of TED talks. Hence, adjectives of judgment are used to express capacity, resolution and veracity in the presentation of scientific data (Martin 2000), which, in turn, helps to construct the researchers’ professional identity (Pérez-Llantada 2024b). Thus, attitudinal adjectives express prominence, intellect and pragmatic functions (see McGrath and Kuteeva 2012 for the case of pure maths research articles), highlighting the similarity of crowdfunding proposals with other knowledge dissemination genres for public understanding and audience engagement in science, such as TED talks (Scotto di Carlo 2014) and citizen science projects (Pérez-Llantada 2021b, 2023). In TED talks, adjectives were used to emphasise the importance of scientists’ contribution to the academic community (Scotto di Carlo 2014) while creating proximity with the audience in a way similar to crowdfunding proposals. In citizen science projects (Pérez-Llantada 2021b, 2023), evaluative adjectives and adverbs together with *I/we* pronouns and static (epistemic/mental) verbs (e.g., *think*, *wonder*) help to create researchers’ identities, constructing competent, credible and trustable selves to appeal to the audience’s *pathos*.

Also, the ‘Discussion’ section, which is a space for microblogging in which to express the backers’ interest in the content of the research, involved evaluation through adjectives that expressed accuracy (e.g., *true*, *right*, *wrong*), quality or emotions. This is a recurrent feature of other digital genres such as academic weblogs (Luzón 2012) and tweets (Villares 2023). These genres used posts to present comments to engage in discussions. The high presence of both positive and negative attitudinal and judgement adjectives in the ‘Discussion’ section indicated that they worked as interactive resources, especially for backers to congratulate researchers profusely or to express interest (and objections) in the projects. Conversely, the most frequent stance particles in the ‘Overview’ and ‘Lab Notes’ sections, that is, modal, attitude, effort verbs and judgement, attitudinal adjectives, mostly evaluated elements within the presentation/adequacy of the research and the budget and expertise of researchers, being therefore used for informative and engagement purposes (i.e., context and significance of the projects, social and scientific impact when backing). This suggests that crowdfunding platforms as Experiment.com are seen as a space for science education and science support. Thus, the majority of the stance words in these sections served to express opinions (attitudes and judgments of value), in line with other popularization genres (e.g., TED talks in Scotto di Carlo 2014 or academic blogs in Luzón 2013), using these linguistic features to indicate affective responses or reactions to the research carried out.

On the other hand, by making judgements and comparisons about protocols and the research process in the ‘Methods’ and ‘Lab Notes’ tabs, writers construct their identity and authority as members of the scientific community while at the same time enhancing the visibility and transparency of their work. Expressing attitudes and opinions in these sections is important for that identity construction. The study findings also suggest that stance adverbials are not common in these projects, although in conversational texts, as in the ‘Discussion’ section, their presence is higher, as it normally occurs in conversation (Conrad and Biber 2000), especially used to mark suggestions, serving to agree with the researchers and their projects. Interestingly, most comments in this section, as well as in the ‘Lab Notes’ section, were positive, contrary to what happens in other digital genres such as microblogs (e.g., *Reddit*) in which popular science is perceived as untrustworthy (Batchelor 2023).

Lastly and more broadly, in response to RQ3 (How does the communicative function of evaluation, as a rhetorical strategy, reflect the communicative purposes and

social action of the genre?), this exploratory study has shown that in crowdfunding proposals online evaluation is important to understand how scientists express their values and opinions about their scientific research in order to reveal themselves as socially situated writers. As seen in the examples, they used stance to construct their identities (in terms of professionalism, importance and transparency) and depict themselves as capable subjects to carry out their research out of the traditional grant circuits and to seek social participation within the scientific community. At the level of rhetorical organization, scientists followed different moves and steps when writing a crowdfunding proposal online in order to present the information about their research to wider lay audiences (Mehlenbacher 2017, 2019) which must be convincing of the importance of the project for the benefit of society. In the same way, at the level of discourse they appear to choose between “rhetorical strategies from a network of linguistic/non-linguistic strategies and end up with their (more or less) personalised versions of this genre” (Askehave and Nielsen 2005: 123).

Aligning with stance studies of other digital genres such as academic blogs, popular science articles, and *Twitter* (e.g. Bondi 2009; Mauranen 2021; Batchelor 2023; Villares 2023), the results of the present study suggest that the functionality of evaluative features is a response to the rhetorical situation underpinning the genre as a social action. As seen in the corpus, the evaluative features in the different rhetorical sections or tabs fulfil several functional goals. Not only do they express the writers’ opinions about their scientific research, but they also serve to construct an identity and build a relationship with potential backers. The inclusion of evaluative features in rhetorical sections such as ‘Meet the Team’ serves to engage the audience, helping them to empathize with the researchers through biographical data rich in attitudinal adjectives, and to persuade them of the campaign’s intent. In addition, by using accessible language, researchers present their updates and steps forward in their project in a detailed and captivating way, engaging backers in the research process. This way, crowdfunding campaigns transform knowledge dissemination into an educational (didactic) reading that makes science accessible and decipherable to the general public. For crowdfunding proposals, an appeal to a large and diversified audience requires accommodating one’s rhetorical efforts to this audience (Mehlenbacher 2019) and that complex audience implies a shift in the use of stance as a strategy to raise funds and educate in science, for example, stating scientific implications in such a way that are also obvious to non-experts, while transmitting an image of

professionalism. It is also worth noting that stance conveying through the use of evaluative grammatical categories in these projects allows to know the scientists' opinion concerning the likelihood of the projects, expressing style stance (Conrad and Biber 2000) in order to contextualise scientific discourse and express the possibilities for carrying out the projects if the necessary funding is raised. The use of stance features in crowdfunding campaigns allows the audience to understand and accept the researchers' perspective, also guiding the audience to accept their claims, and thus it will be more prone to act on the researchers' call-to-action. There was evidence in the corpus data that the use of attitudinal verbs and adjectives alongside positive judgment adjectives helped crowdfunding writers to emphasise the noteworthiness of their projects' content and "the positive aspects connected to the world of science" (Scotto di Carlo 2014: 214).

This exploratory work is, however, limited by the small size of the corpus, which has conditioned the analysis at the section/tab level. On the other hand, the lack of inferential statistical analysis to corroborate and/or refute the differences found between grammatical categories and between the use of stance markers in the different sections of these proposals has also been a limitation. Thus, this study leaves many analytical gaps to be filled in future research using inferential statistics. Although the finding suggest that language features to express evaluation are used to realize specific rhetorical functions in the proposals (e.g., inform, greet, express admiration) it would be important to explore the use of the different grammatical categories of stance in the projects according to variables such as the scientists' expertise in the field (i.e., whether they relate to expertise in the field, that is junior or senior researchers) and expertise and prior knowledge in composing science-related texts. Ethnomethodologically informed studies (i.e., interviews) could shed new light on the degree of consciousness or awareness when using these stance grammar categories in line with other studies about attitudinal and epistemic stance (e.g., Martin and White 2005; Hyland 2005). It is important to know the level of awareness of the use of evaluative language in composing these proposals and their rhetorical effects. Knowledge of composing strategies across genres, modes and media can support researchers when they want or need to compose a project proposal (e.g., the use of language for the creation of a scientific community and collaboration, the use of language for identity construction, the use of language for social commitment and tolerance).

REFERENCES

- Askehave, Inger and Anne E. Nielsen. 2005. Digital genres: A challenge to traditional genre theory. *Information Technology and People* 18/2: 120–141.
- Batchelor, Jordan. 2023. Just another clickbait title: A corpus-driven investigation of negative attitudes toward science on Reddit. *Public Understanding of Science* 32/5: 580–595.
- Biber, Douglas. 2004. Historical patterns for the grammatical marking of stance: A cross-register comparison. *Journal of Historical Pragmatics* 5: 107–136.
- Biber, Douglas. 2006. *University Language: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.
- Biber, Douglas and Edward Finegan. 1989. Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text* 9: 93–124.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Pearson Education Ltd.
- Bondi, Marina. 2009. Perspective and position in museum websites. In Sara Radighieri and Paul Tucker eds. *Point of View. Description and Evaluation across Discourses*. Roma: Officina Edizioni, 113–127.
- Brezina, Vaclav and William Platt. 2023. #LancsBox v. 6.x. [software package].
- Camiciotti, Gabriella del Lungo and Elena Tognini-Bonelli eds. 2004. *Academic Discourse: New Insights into Evaluation*. Bern: Peter Lang.
- Conrad, Susan and Douglas Biber. 2000. Adverbial marking of stance in speech and writing. In Susan Hunston and Geoffrey Thompson eds. *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: Oxford University Press, 56–73.
- Dave, Kushal, Steve Lawrence and David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web (WWW'03)*. New York: Association for Computing Machinery: 519–528. <https://dl.acm.org/doi/10.1145/775152.775226>
- De Waard, Anita and Henk Pander Maat. 2012. Epistemic modality and knowledge attribution in scientific discourse: A taxonomy of types and overview of features. In Haizhou Li, Chin-Yew Lin, Miles Osborne, Gary Geunbae Lee and Jong C. Park eds. *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*. Jeju Island, Korea: Association for Computational Linguistics, 47–55. <https://aclanthology.org/W12-4306.pdf>
- Englebretson, Robert. 2007. Stancetaking in discourse: An introduction. In Robert Englebretson ed. *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*. Amsterdam: John Benjamins, 1–25.
- Follett, Ria and Vladimir Strezov. 2015. An analysis of citizen science based research: Usage and publication patterns. *PLoS ONE* 10/11: e0143687. <https://doi.org/10.1371/journal.pone.0143687>
- Goźdz-Roszkowski, Stanisław and Susan Hunston. 2016. Corpora and beyond. Investigating evaluation in discourse: Introduction to the special issue on corpus approaches to evaluation. *Corpora* 11/2: 131–141.

- Hunston, Susan. 1989. *Evaluation in Experimental Research Articles*. Birmingham: University of Birmingham dissertation.
- Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
- Hunston, Susan. 2007. Semantic prosody revisited. *International Journal of Corpus Linguistics* 12: 249–268.
- Hunston, Susan. 2011. *Corpus Approaches to Evaluation: Phraseology and Evaluative Language*. London: Routledge.
- Hunston, Susan and John Sinclair. 2000. Towards a local grammar of evaluation. In Susan Hunston and Geoffrey Thompson eds. *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: Oxford University Press, 74–101.
- Hyland, Ken. 1999. Academic attribution: Citation and the construction of disciplinary knowledge. *Applied Linguistics* 20: 341–367.
- Hyland, Ken. 2002. Activity and evaluation: Reporting practices in academic writing. In John Flowerdew ed. *Academic Discourse*. London: Longman, 115–130.
- Hyland, Ken. 2005. Stance and engagement: A model of interaction in academic discourse. *Discourse Studies* 7/2: 173–192.
- Hyland, Ken and Feng Jiang. 2017. Is academic writing becoming more informal? *English for Specific Purposes* 45: 40–51. <https://doi.org/10.1016/j.esp.2016.09.001>
- Jiang, Feng and Ken Hyland. 2015. ‘The fact that’: Stance nouns in disciplinary writing. *Discourse Studies* 17/5: 529–550.
- Luzón, María José. 2012. ‘Your argument is wrong’: A contribution to the study of evaluation in academic weblogs. *Text & Talk* 32/2: 145–165.
- Luzón, María José. 2013. Public communication of science in blogs: Recontextualizing scientific discourse for a diversified audience. *Written Communication* 30: 428–457.
- Luzón, María José. 2023. Multimodal practices of research groups in Twitter: An analysis of stance and engagement. *English for Specific Purposes* 70: 17–32.
- Luzón, María José and Sofía Albero-Posac. 2020. ‘Had a lovely week at #conference2018’: An analysis of interaction through conference tweets. *RELIC Journal* 51/1: 33–51.
- Luzón, María José and Carmen Pérez-Llantada. 2022. *Digital Genres in Academic Knowledge Production and Communication: Perspectives and Practices*. Bristol: Multilingual Matters.
- Lyons, John. 1981. *Language and Linguistics: An Introduction*. Cambridge: Cambridge University Press.
- Martin, James Robert. 2000. Close reading: Functional linguistics as a tool for critical discourse analysis. In Len Unsworth ed. *Researching Language in Schools and Communities*. London: Cassell, 275–302.
- Martin, James Robert and Peter R. White. 2005. *The Language of Evaluation*. New York: Palgrave Macmillan.
- Mauranen, Anna. 2002. ‘A good question’: Expressing evaluation in academic speech. In Giuseppina Cortese and Peter Riley eds. *Domain-specific English: Textual Practices Across Communities and Classrooms*. Bern: Peter Lang, 115–140.
- Mauranen, Anna. 2003. The corpus of English as lingua franca in academic settings. *TESOL Quarterly* 37/3: 513–527.
- Mauranen, Anna. 2021. ‘Gonna write about it on my blog too’: Metadiscourse in research blog discussions. In Larissa D’Angelo, Anna Mauranen and Stefania Maci eds. *Metadiscourse in Digital Communication: New Research, Approaches, and Methodologies*. London: Palgrave Macmillan, 11–35.

- McGrath, Lisa and Maria Kuteeva. 2012. Stance and engagement in pure mathematics research articles: Linking discourse features to disciplinary practices. *English for Specific Purposes* 31/3: 161–173.
- Mehlenbacher, Ashley Rose. 2017. Crowdfunding science: Exigencies and strategies in an emerging genre of science communication. *Technical Communication Quarterly* 26/2: 127–144.
- Mehlenbacher, Ashley Rose. 2019. *Science Communication Online: Engaging Experts and Publics on the Internet*. Ohio: The Ohio State University Press.
- Millar, Neil, Bojan Batalo and Brian Budgell. 2023. Trends in the use of promotional language (hype) in abstracts of successful national institutes of health grant applications. *JAMA Netw Open* 5/8: 1985–2020.
- Nasukawa, Tetsuya and Jeonghee Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing. In John H. Gennari, Bruce W. Porter and Yolanda Gil eds. *Proceedings of the 2nd International Conference on Knowledge Capture*. Sanibel Island, FL, USA: Association for Computing Machinery, 70–77. <http://dx.doi.org/10.1145/945645.945658>
- Pérez-Llantada, Carmen. 2021a. Grammar features and discourse style in digital genres: The case of science-focused crowdfunding projects. *Revista Signos* 54/105: 73–96.
- Pérez-Llantada, Carmen. 2021b. *Research Genres Across Languages*. Cambridge: Cambridge University Press.
- Pérez-Llantada, Carmen. 2022. Online data articles: The language of intersubjective stance in a rhetorical hybrid. *Language Communication* 39/3: 400–425.
- Pérez-Llantada, Carmen. 2024a. Approaching digital genre composing through reflective pedagogical praxis. *Journal of English for Academic Purposes* 68: article 101349.
- Pérez-Llantada, Carmen. 2024b. Identity construction in digital communication for public engagement in science. *Discourse Studies* 27/1: 128–145. <https://doi.org/10.1177/14614456241255267>
- Pérez-Llantada, Carmen and María José Luzón. 2023. *Genre Networks. Intersemiotic Relations in Digital Science Communication*. New York: Routledge.
- Rahimpour, Sepideh. 2014. Blogs: A resource of online interactions to develop stance-taking. *Procedia. Social and Behavioral Sciences* 98: 1502–1507.
- Sancho-Guinda, Carmen and Ken Hyland eds. 2012. *Stance and Voice in Written Academic Genres*. New York: Palgrave-McMillan.
- Scotto di Carlo, Giuseppina. 2014. The role of proximity in online popularisations: The case of TED talks. *Discourse Studies* 16/5: 591–606.
- Shaw, Philip. 2003. Evaluation and promotion across languages. *Journal of English for Academic Purposes* 2: 343–357.
- Smith, Alison. 2015. ‘Wow, I didn’t know that before; thank you’: How scientists use Twitter for public engagement. *Journal of Promotional Communications* 3/3: 320–339.
- Swales, John M. 2004. *Research Genres. Explorations and Applications*. Cambridge: Cambridge University Press.
- Swales, John M. and Amy Burke. 2003. ‘It’s really fascinating work’: Differences in evaluative adjectives across academic registers. In Pepi Leistyna and Charles F. Meyer eds. *Corpus Analysis. Language and Structure and Language Use*. Amsterdam/New York: Editions Rodopi, 1–18.
- Thompson, Geoffrey. 2014. *Introducing Functional Grammar* (third edition). London: Routledge.

- Thompson, Geoffrey and Susan Hunston. 2000. Evaluation: An introduction. In Geoffrey Thompson and Susan Hunston eds. *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: Oxford University Press, 1–27.
- Turney, Peter. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia. July 2002: 417–424.
- Vela-Rodrigo, Alberto Á. 2023. A lexical bundle analysis of art-related crowdfunding projects. *Ibérica, Journal of the European Association of Languages for Specific Purposes* 46: 321–349.
- Vela-Rodrigo, Alberto Á. 2025 (in press). A case study for the analysis of the rhetoric of crowdfunding communication. *LFE: Revista de Lenguas para Fines Específicos*, 31.
- Villares, Rosana. 2023. Twitter conference presentations: A rhetorical and semiotic analysis of an emerging digital genre. *ELIA: Estudios de Lingüística Inglesa Aplicada* 22: 125–167.
- Zou, Hang and Ken Hyland. 2019. Reworking research: Interactions in academic articles and blogs. *Discourse Studies* 21/6: 713–733.

Corresponding author

Alberto Ángel Vela-Rodrigo
 University of Zaragoza
 Faculty of Social and Human Sciences
 Department of English and German Studies
 Campus in Teruel
 C/ Atarazanas, 4
 44003, Teruel
 Spain
 E-mail: vela@unizar.es

received: October 2024

accepted: April 2025

Spanish EFL learners' use of contrastive linking adverbials across three CEFR levels and gender

Carmen Maíz-Arévalo
Complutense University of Madrid / Spain

Abstract – The *Common European Framework of Reference for Languages* (2001) and its *Companion Volume* (2020) emphasise the importance of linking expressions for pragmatic competence. Research on contrastive linking has long attracted scholarly interest; however (pseudo)longitudinal studies across different levels or whether gender may affect learners' written production in this respect have been neglected. This study aims to address this gap by analyzing how Spanish English as a Foreign Language (EFL) learners at different levels express contrast and whether gender impacts their use of concessive expressions. Surprisingly, lower-level (B1) users show a wide range of expressions similar to higher-level users, while those at B2 levels tend to avoid 'risky' options. Interestingly, gender does not significantly influence learners' use of connectors in this corpus, contradicting earlier findings that suggested female learners use more connectors than males.

Keywords – Spanish EFL learners; CEFR/CV descriptors; contrastive connectors; gender.

1. INTRODUCTION

According to the *Common European Framework of Reference for Languages* (CEFR) (Council of Europe 2001) and its *Companion Volume* (CV) (Council of Europe 2020), learners of a foreign language should acquire three communicative language competences: linguistic competence, sociolinguistic competence, and pragmatic competence. Pragmatic competence encompasses thematic development, which is concerned with “the ability to design texts, including generic aspects like Thematic development and Coherence and cohesion” (CV 2020: 139). Coherence and cohesion refer to the way “in which the separate elements of a text are interwoven into a coherent whole by exploiting linguistic devices such as referencing, substitution, ellipsis and other forms of textual cohesion, plus logical and temporal connectors and other forms of discourse markers” (*ibid.*, 142).



The study of linking adverbials cohesive devices by EFL learners has attracted much scholarly attention for decades, with a special focus on the use of adverbs (e.g. Gilquin and Paquot 2008; Yilmaz and Dikilitas 2017, among many others), often in contrast with native speakers/writers' use, a common methodological approach being that of contrastive interlanguage analysis (Granger 2015). Research has shown that learners do not appear to find it too problematic to employ time and place as logical connectors. However, this does not seem to be the case with concessive connectors such as *however* or *yet*, which has been shown to be harder to master for EFL learners (Celce-Murcia and Larsen-Freeman 1999).

In the case of L1 Spanish EFL learners, the adverbial expression of contrast has proven to be particularly difficult, with learners both transferring from their L1 and overusing the same devices (Pérez-Paredes *et al.* 2012; Carrió-Pastor 2013; Jiménez Catalán and Ojeda Alba 2014; Mora Díaz and Gómez Orjuela 2021; Faya-Cerqueiro and Martín Macho-Harrison 2022, among others). Nevertheless, these studies have focused on the analysis of different cohesive devices at specific levels rather than contrasting the use of such devices across different levels and in connection with the sociological variable of gender. The aim of this paper is to redress this imbalance by analysing contrastive connectors in a learner corpus across the levels B1, B2, and C1 to find out to what extent (if any) linguistic proficiency and gender influences the way learners employ these connectors in their opinion essay writings.

There are two reasons for choosing these three levels (B1, B2, and C1). On the one hand, according to the CEFR and the CV, learners should be able to use cohesive devices from level B1 onwards, augmenting their range proportionally to their level. The description is rather vague as it is not clear what this 'range' actually means. Thus, the B1 descriptor claims that students "can introduce a counter-argument in a simple discursive text (e.g. with *however*)," while the B2 descriptor states that students "can produce text that is generally well-organised and coherent, using *a range of linking words and cohesive devices*" (my emphasis). Moving up, at the C1 levels learners are expected to be able to "produce well-organised, coherent text, using a variety of cohesive devices." On the other hand, these are the three levels under scrutiny in the FineDesc Project,¹ within which the present research is encompassed.

¹ This research has been funded by the Spanish Ministry of Science, Innovation and Universities and the National Research Agency. Grant PID2020-117041GA-I00, funded MICIU/AEI/10.13039/501100011033

The rest of this paper is structured as follows. Section 2 has been divided into two subsections. Section 2.1 describes the theoretical background, with special attention to previous taxonomies of contrast devices and the rationale behind the choice of the linking adverbials chosen in the present study. Section 2.2 reviews previous studies on the use of contrast devices by non-native EFL learners, with a special emphasis on Spanish learners. Section 3 describes the methodology, including a detailed explanation of the corpus and the tools employed in the present analysis. Section 4 discusses the findings before presenting the conclusions in Section 5, together with some pointers to future research.

2. LINKING ADVERBIALS IN ENGLISH ACADEMIC WRITING

2.1. *Linking adverbials as cohesive devices*

A review of the theoretical background reveals a lack of consensus regarding nomenclature. Thus, Quirk *et al.* (1985: 440) term words like *however* or *nevertheless* ‘conjuncts’; that is, words that “express the speaker’s assessment of the relation between two linguistic units.” On the other hand, Huddleston and Pullum (2002) prefer the term ‘connective adjuncts’, while Blakemore (2006: 221) uses the term ‘discourse markers’, which is defined as “generally used to refer to a syntactically heterogeneous class of expressions which are distinguished by their function in discourse and the kind of meaning they encode.” Cowan (2008: 615) refers to ‘discourse connectors’ as “words and phrases that, typically, connect information in one sentence to information in previous sentences.” Biber *et al.* (2021: 137) used the term ‘linking adverbials’, which are said to “express the type of connection between clauses.”

Despite the differences in terminology, all authors establish a clear distinction between linking adverbials and coordinating and subordinating conjunctions. In fact, Biber *et al.* (2021) specify that coordinators are closely related to linking adverbials, although there are three main syntactic differences:

- i) The position of the coordinator is fixed within the clause boundary, while linking adverbials are more flexible. E.g., *They **nevertheless** decided to go camping* versus **They **but** decided to go camping*.
- ii) Linking adverbials may be preceded by coordinators, while the latter are mutually exclusive. E.g., *And **nevertheless**, they decided to go camping* versus **And **but** they decided to go camping*.

- iii) Linking adverbials are often marked by commas, while coordinators are not. E.g., *Nevertheless, they decided to go camping* versus **But, they decided to go camping*.

Furthermore, conjunctions are often employed to create shorter and more concise sentences while conjunctive adverbs (or linking adverbials) tend to render more complex sentences. In the present study, and for the sake of space, I will be focusing on linking adverbials of contrast in the written expression of EFL Spanish learners.

Linking adverbials are classified by Quirk *et al.* (1985: 634) into seven categories, which are further subdivided into ten subcategories. Among these, the contrastive category includes four subtypes: reformulatory (*better*), replacive (*on the other hand*), antithetic (*conversely*), and concessive (*however*). These four subtypes, however, are not always clear-cut.

More recently, Biber *et al.* (2021) divide linking adverbials into six categories, one of which is contrast/concession. Regarding their form, they can typically be realised by single adverbs such as *nevertheless* or *still*, adverb phrases like *even so*, and prepositional phrases such as *by contrast* or *on the other hand*. For the present study, I will be following this classification given its more comprehensive definition of the contrastive-concessive subcategory and its syntactic characteristics. Thus, I will focus on the analysis of the following two groups of concessive linking adverbials: (i) single adverbs *however*, *nonetheless*, *nevertheless*, *yet*, and *still*; and (ii) prepositional phrases *on the other hand*, *by contrast*, and *in contrast*.

2.2. Learners' use of linking adverbials: Previous research

The analysis of written compositions by learners of English as a foreign or second language (EFL/ESL) has been a subject of scholarly interest for decades, giving rise to a vast amount of research on learners of many different linguacultural backgrounds and L1s such as Arabic (Modhish 2012; Appel 2020; Ahmed *et al.* 2023), Chinese (Lee 2020; Zhang 2021), French (Granger and Tyson 1996), Hungarian (Tankó 2008), Iranian (Hosseinpour and Pour 2022), Japanese (Narita *et al.* 2004), Korean (Lee 2013; Ha 2016; Yoon 2019), Norwegian (Hasselgren 1994), or Swedish (Altenberg and Tapper 1998), among many others.

Despite the wide variety of L1s, findings tend to consistently reveal three tendencies (not always mutually exclusive): (i) non-native speakers often exhibit patterns of overuse in contrast to native users (what has been termed the ‘overuse hypothesis’); (ii) EFL learners misuse certain connectors due to transfer from their native language (see Granger and Tyson 1996); and (iii) learners tend to display a limited range of devices in contrast to their native counterparts.

The overuse tendency has, however, been widely reported, often in combination with the lack of variation. For example, in a recent study of over 180 upper-intermediate and advanced Iranian EFL students, Hosseinpour and Pour (2022) showed that the students lacked variation and overused the contrastive markers *nevertheless*, *in contrast*, *on the other hand*, or *on the contrary*. This tendency has also been reported among Arabic EFL students. Thus, Ahmed *et al.* (2023) have shown that even high proficiency students rely on *but* as their preferred adversative connector. These results are in line with previous research on Arabic EFL learners, which additionally proved that users tended to employ both *but* and *however*, overusing both connectors and displaying a limited range of devices (Modhish 2012; Appel 2020).

The use of linking adverbials by Chinese learners has also received recent attention. For example, Zhang (2021) provides a comprehensive analysis of adversative and contrastive expressions (both paratactic and hypotactic) by EFL Chinese students, whose usage he contrasts with that of native writers. The author shows that, in contrast to native speakers, Chinese students overuse *on the other hand* to express contrast but underuse other expressions like *however* or *nevertheless* in contrast to native speakers. Other expressions like *yet* or *still* with a contrastive meaning are also neglected by his Chinese participants. In conclusion, his study shares with previous studies the fact that non-native learners often overuse a small set of expressions in contrast to the wider range of connectors employed by their native counterparts.

Other Asian speakers have also received scholarly attention. For example, Lee (2013) showed that Korean students significantly overused contrastive markers compared to native writers, who preferred to make implicit statements of contrast. Korean learners lacked variation, with a heavy use of *but*, and an underuse of other markers like *yet* and *however*. These results, interestingly, did not seem to improve when students’ level was higher. Park (2013) also contrasted Korean EFL and native writers, with the former having either a low or an advanced level. She found that both groups showed a preference

for *but* and *however*, with the non-native writers resorting to more rigid syntactic position than their native counterparts (see also Ha 2016; Yoon 2019).

The use of linking devices by Spanish EFL learners has also been widely studied. For example, Lahuerta Martínez (2002) found that these learners tended to overuse a limited range of three concessive connectors: *but*, *although*, and *however*, although this study is limited since it focuses on the written compositions of only seven participants. Studies based on larger corpora have shown similar results. For example, Rica Peromingo (2012) focused on B1 and B2 Spanish EFL learners, showing that these students used more contrastive connectors than their native peers and that they prefer multi-word units like *on the other hand* rather than single contrastive adverbs. This overuse hypothesis was also proved by Navarro Gil and Roquet Pugès (2020), who studied a corpus of 50 argumentative essays, written by Spanish EFL learners, among others. These were later contrasted with a native corpus to study the use of adversative linking adverbials, which resulted in non-native speakers using and placing them more often in sentence-initial position. According to the authors, these learners often overuse adversative linking adverbials and tend to employ them in rather fixed positions, often leading to punctuation issues, even if they have an advanced level.

Language transfer has also been reported among Spanish EFL learners, as shown by Neff-van Aertselaer and Dafouz-Milne (2008). After comparing two corpora (one made up of argumentative essays by Spanish EFL university students (with 194,845 tokens), and the other written by British and American students (149,790 tokens)), the authors concluded that:

Regarding textual metadiscourse [...] this study shows that both novice groups under-use logical markers (e.g. *and*, *moreover*, *but*, *however*, *nevertheless*, *yet*; etc.) but over-use sequencers (*first*, *second*, *next*, etc.), and the Spanish EFL writers much more so than American university students. This feature may reflect Spanish rhetorical conventions, which favour a progressive argumentation strategy, building up evidence of the same type, clause by clause, while English argumentation strategies prefer setting out the major premise at the beginning and then offering a balanced consideration of the pros and cons. (Neff-van Aertselaer and Dafouz-Milne 2008: 98)

For other authors, however, the fact that non-native writers employ connectors differently from native writers is not so much due to negative language transfer from the formers' L1 but rather a result of their linguistic background, which leads to "making

different rhetorical choices to construct identity while maintaining text coherence in academic discourse” (Carrió-Pastor 2013: 193). Interestingly, in her contrastive study between two academic corpora (by native and non-native writers), the author did not find any significant difference in the frequency of use of contrastive connectors like *however*, which were used similarly by both groups. She found out that there was a lack of variety in the range of connectors employed by the non-native writers, in line with previous research.

Together with learners’ level, another variable that has long attracted scholars’ attention is gender. In the field of discourse markers, the work by Tavakoli and Karimnia (2017) is worth mentioning; they investigated the Iranian EFL learners’ use of discourse markers in connection with gender and their results show that female users tend to employ more discourse markers than their male counterparts. In a similar line, Alqahtani and Abdelhalim (2020) studied 60 academic essays written by EFL university students (30 by female and 30 by male students), with a focus on metadiscourse markers. They revealed a statistically significant difference between male and female students in using some interactive markers such as transitions, frame markers, and code glosses, with female students surpassing their male counterparts.

In the case of Spanish EFL learners, the study of gender has focused on lexical choices (Díez Prados 2010), the expression of emotion (Pérez-García and Sánchez 2020), or the use of the L1 during interaction (Azkarai 2015). Results seem not to be conclusive, with some of the previous authors reporting differences between genders. Nevertheless, in a more recent study on the written expression of EFL Mexican students, Núñez Mercado (2022) has found no significant difference in terms of discourse markers by the two genders. To the best of my knowledge, the study of contrastive connectors across levels and in connection with gender has not been tackled in the study of Spanish EFL learners.

3. DATA AND METHODS

The present paper intends to redress the previously mentioned imbalance by aiming to answer the following research questions:

RQ1. To what extent (if any) is the use of contrastive linking adverbials affected by the learners’ CEFR level? In other words, to what degree do higher-level students

employ a wider range of linking adverbials in contrast to lower-level students? More specifically, I will be analysing two groups of contrastive linking adverbials: (i) the single adverbs *however*, *nonetheless*, *nevertheless*, *yet*, and *still*; and (ii) the prepositional phrases *on the other hand*, *by contrast*, and *in contrast*.

RQ2. To what extent (if any) does the learners' gender affect their use of these contrastive linking adverbials (i.e., frequency of use and syntactic variety of use)?

In line with previous research (see Neff-van Aertselaer and Dafouz-Milne 2008; Rica Peromingo 2012; Carrió-Pastor 2013; Navarro Gil and Roquet Pugès 2020), it is hypothesised that the higher the learners' level, the wider range of concessive expressions they will be able to employ and the more varied their discursive position. In other words, it is expected that B1 learners will more often resort to *however* in sentence-initial position while C1 learners are hypothesised to use other concessive adverbials such as *nonetheless* and deploy a wider range of discursive positions in the sentence (initial, middle, and final). B2 learners are predicted to stand in between the previous and following levels. Regarding gender, and according to prior research (see Tavakoli and Karimnia 2017; Alqahtani and Abdelhalim 2020), it is hypothesised that female learners will use contrastive connectors more frequently than their male counterparts while no differences are expected with regard to variety.

The data set employed in the present study is part of the FineDesc learner-based corpus (gathered by the FineDesc project²). This corpus includes the written compositions of EFL students who take official accreditation exams to certify their level in different language centres all over Spain. This factor is important as the candidates are taking part in a high stakes exam and often wish to do their best to pass and obtain their certification. All the candidates were asked to fill in a consent form, but to preserve their privacy all the exams have been transcribed and anonymised, numbered and coded according to the gender³ reported by the candidates themselves. Furthermore, to make sure the level of the students corresponds to the three levels under scrutiny (i.e., B1, B2, and C1), all the essays included in the corpus (both the general corpus and the sub-set here analysed) belong to students who passed the exam, hence obtaining a certification of their level. Finally, all

² <https://web.ujaen.es/investiga/finedesc/index.php>

³ Participants were provided with the male, female, non-binary, and rather not say options when asked about their gender. All the participants in the present dataset reported to be either male or female.

the data thus compiled have been approved by the main researcher's university ethical committee.

For the purposes of the current paper, all the selected writings belong to the opinion essay genre on different topics such as rural versus city life, the pros and cons of technology, or the advantages and disadvantages of bilingual education, among others. This choice was determined by the expectation that it is in this genre that learners are more likely to employ contrastive expressions. Table 1 sums up the gender and number of participants involved in each level. Table 2 presents the number of words in each of the three levels. The corpus is thus a convenience sample as it is informed by the candidates who passed each of the levels of the accreditation exam. Neither the number of candidates per level nor the gender could be determined by the researcher beforehand.

CEFR Level	B1 level		B2 level		C1 level	
Gender	Female	Male	Female	Male	Female	Male
	41	41	70	70	25	28
Total	82		140		53	

Table 1: Corpus description (levels and participants)

Level	B1 level	B2 level	C1 level	TOTAL
Words	12,202	30,846	14,894	57,942

Table 2: Corpus description (number of words per level)

Though admitting that the dataset is limited, it is possible to approach it from a mixed-methods perspective (see Ghadessy *et al.* 2008). The quantitative part has been carried out with the aid of Sketch Engine, where the corpus was uploaded. However, given that some of the students' linguistic productions were not always automatically retrieved, a manual retrieval was included. For example, *on the other hand* may also appear as *on another hand*, which made the manual retrieval essential. Likewise, adverbs like *yet* or *still* could not be retrieved using exclusively automatic methods as they are polysemous and, besides contrast, they may also function as time relationship adverbials (Quirk *et al.* 1985: 194). More specifically, I have focused on the frequency, discursive position, and use of the following contrastive linking adverbials: the adverbs *however*, *nonetheless*, *nevertheless*, *yet*, and *still* and the prepositional phrases *on the other hand*, *in contrast*,

and *by contrast*. Excel was used to normalise frequencies per thousand words (ptw) so as to make the subcorpora comparable, given their different sizes.⁴

4. RESULTS AND ANALYSIS

For the sake of clarity, Section 4 has been divided into three sub-sections, the first two of which align with the corresponding research questions. Thus, Section 4.1 focuses on the use of contrastive linking adverbials across the three levels under study (B1, B2, and C1), while Section 4.2 focuses on the contrast between both genders. Finally, Section 4.3 includes a summarising general discussion and links results with the pertinent descriptors of the CEFR and Companion Volume (2020).

4.1. Contrastive linking adverbials across levels

Table 3 offers an overview of the frequency of all the adverbial links under scrutiny across the three levels. The left column shows the raw number of occurrences, and the right column presents the normalised frequency per thousand words (ptw). *Still* and *by contrast* produced no hits. In the case of *still*, this was after disambiguation.

	B1		B2		C1	
Linking adverbial	Freq. (raw)	Freq. (ptw)	Freq. (raw)	Freq. (ptw)	Freq. (raw)	Freq. (ptw)
<i>However</i>	20	1.639	51	1.653	20	1.342
<i>Nevertheless</i>	3	0.245	13	0.421	8	0.537
<i>Nonetheless</i>	1	0.081	0	0	3	0.201
<i>Yet</i>	1	0.081	0	0	3	0.201
<i>In contrast</i>	1	0.081	0	0	1	0.067
<i>On the other hand</i> ⁵	32	2.622	69	2.236	17	1.141

Table 3: Overview of frequency of use

⁴ Given the size of the dataset, it was decided to normalise per thousand words rather than per million words.

⁵ For the purposes of the current research, only the phrase expressing contrast has been considered, including also grammatically wrong expressions such as *in the other hand* or *on another hand*, which were manually retrieved but also quantified and included in this normalised frequency. Future research might zero in on its other uses such as expressing an alternative.

As can be seen, B1 and B2 learners favour the use of *on the other hand* as their most frequent way to express contrast, followed by *however*. C1 students follow a similar pattern, with a preference for these two adverbial linkers, but they opt for the single adverb *however* slightly more frequently than *on the other hand*. The third most frequent linking adverbial across the three levels is *nevertheless*. While these results are to be expected in line with previous research, it is remarkable that B2 students are those that deploy the lowest degree of variation, limiting their options to the three adverbial linkers already mentioned; in contrast, B1 learners show a slightly more varied range, shadowing the patterns of C1 students, albeit at a lower frequency. As expected, C1 learners are those with a wider range of variety. The outcome of the chi-square test was significant ($p=0.042$).

In the following paragraphs, I will focus on each of these expressions individually according to its level of frequency in the dataset, namely, *on the other hand*, *however*, *nevertheless*, *nonetheless*, *yet*, and *in contrast*. The remaining two linking adverbials (i.e., *still* and *by contrast*) have no tokens in any of the three levels, but I will try to explain the plausible reasons for this absence.

As shown in Table 3, there is a progressive decrease in the use of *on the other hand* across the three levels, with B1 students using it also more frequently than *however*. The use seems to decrease by more than half when students reach C1 level (from 2.622 ptw in B1 to 1.141 ptw in C1). A plausible reason for such a drop might be the influence of instruction. For example, the *Cambridge Writing Guide for C1* (Porras Wadley 2022) recommends using more formal linking phrases such as *nevertheless* or *nonetheless*. Instructors, as a result, may play an influential role in learners' choices. For example, Schenck (2020) has shown that these differences between non-native and English native writers may be influenced by their instruction. In his study, he contrasted the writings of EFL Korean students receiving instruction from native-speaking teachers and those who were taught by non-native ones. His results show that students whose instruction was carried out by native teachers were able to express more nuanced opinions and employed a wider range of devices. His study, however, presents an imbalanced sample, as the group taught by native speakers amounts to 59 participants in contrast to 19 in the group that was not.

Furthermore, it is worth noting that only students at the highest level (C1) succeed in writing the connector correctly, while students at B1 occasionally make mistakes with

the accompanying preposition, using prepositions like *in* and even *by*, as in examples⁶ (1) and (2), or employing *another hand*, as in (3):

- (1) **In the other hand**, there are a lot of pitfalls. [9281F]
- (2) **By the other hand**, it could make you lower your grades because of the lack of time for studying. [9299F]
- (3) **In another hand** the game can be bad too if you use it bad the people stay all day sitting at home in from of their computers also they don't practise any sport and they don't do exercise, this can beeing development some disease. [6257F]

These errors, however, amount to 15.6 per cent of the B1 sub-corpus, which may indicate that most students learn the expression as a bundle. More remarkable is that students at B2 level still make the same mistake, although the wrong preposition is always *in*, as in examples (4) and (5) by a male and female learner, respectively:

- (4) **In the other hand**, we can't forget how important englis is now a days. [70106M]
- (5) **In the other hand**, we should keep in mind that not all students have the same learning skills. [10515F]

Although these cases amount to 13 per cent of the sub-corpus (nine out of 69 tokens), it shows that for these learners, it might have become a fossilised error. This connector, albeit used slightly less frequently at this level (0.20% of the sub-corpus) is still rather common among these learners to express contrast, while C1 students show a clear tendency to avoid it, maybe as a result of the aforementioned influence of instruction. These results are in line with Rica Peromingo (2012), whose analysis of Spanish EFL students' written compositions also revealed a preference for *on the other hand* among B1 and B2 learners.

As can be observed in Table 3, *however* is the linking adverbial most favoured by C1 learners (1.342 ptw), although both B1 and B2 levels also employ it frequently (1.639 ptw and 1.653 ptw, respectively). However, there are discursive differences in its use. Thus, when exploring concordances, it is observed that at B1 level, students invariably

⁶ Each example is reproduced as it appears in the original text, followed by a number indicating the number of student and their self-reported gender. Thus, F stands for female and M for male. The word under scrutiny is in bold for the sake of clarity.

place *however* at the beginning of the sentence, after a period and followed by a comma, as in examples (6) and (7) by a female and male learner, respectively:

- (6) will be shown among students of this kind. They'll also learn time management, decision making and planning skills. **However**, it should be taken into consideration that working while studying could bring lack of motivation, cognitive fatigue [...] [9276F]
- (7) As students, we have to work hard to pass the exams and take as much knowledge as it its possible. **However**, everybody knows that the university studetns have a lot of free time.

At this level, students do not seem to be aware of the fact that this adverb can also be placed in middle- and final-sentence position. This could be a result of instruction, as this is the most frequent position exemplified by model texts in didactic materials, which seems to 'prime' students (see Leedham and Cai 2013).

When compared with B2-level students, there is a small attempt at placing the adverb in mid-sentence position (seven tokens out of 51), even if this is mostly done incorrectly and there are punctuation errors, as in examples (8) and (9) by a female and male learner, respectively. There is, however, one case where the student uses it appropriately, in example (10):

- (8) people of other country, improve your CV (...) In short I will say that is very benefit study in a bilingual education **however** also have the disadvantage that some students do not learn the subject properly. [70100F]
- (9) is true that language is a very important addition on your CV. In my opinion, I think bilingual education could be great, **however** if I could choose, I will only study in other language the last year of my degree. [70140M]
- (10) Some people think that schools in our country should be bilingual, others, **however**, think that this type of education is not necessary for their kids. [70119F]

In any case, and despite the errors, students at this level seem to be aware of the possibility to place the adverb in a position other than the beginning of the sentence, even if this still remains the most favoured option (with 44 out of 51 tokens). Interestingly, users at C1 return to what could arguably be described as a more conservative use, as they place all the tokens in initial position.

The use of *nevertheless* increases with proficiency, with C1 learners leading the way (0.537 ptw), followed by B2 (0.421 ptw) and B1 learners (0.245 ptw). Independently

of the level, however, learners always place *nevertheless* in the initial position, followed by a comma, as in examples (11) to (13), corresponding to B1, B2, and C1 level, respectively:

- (11) **Nevertheless**, we have to be aware of the fanger that spending too much time playing can cause. [6258F]
- (12) **Nevertheless**, for some people could be a disadvantage not to know other idioms [10150F]
- (13) **Nevertheless**, some advocates against this conservation argue that keeping these huge ruins suppose the possibility of exposing the population to a real danger, on the grounds of the careless state of this 'old ladies'. [900016M]

Interestingly, the three adverbial linkers already commented in the previous paragraphs (i.e., *on the other hand*, *however*, and *nevertheless*) are those to which all students resort, especially those at B2 level, since these are the only three options they employ throughout the dataset. This makes both B1 and C1 learners' writings more varied in terms of contrastive adverbials. While this tendency is difficult to explain, a plausible reason might be that B2 students are less willing to take risks in order to ensure passing the exam, hence obtaining the corresponding certification. This specific instructional context (i.e., a high-stakes exam) might be rendering these students particularly unwilling to take risks, in line with Brown (2001: 63). According to this author, "many instructional contexts around the world do not encourage risk-taking; instead, they encourage correctness, right answers, and withhold 'guesses' until one is sure to be correct." On the other hand, C1 students' linguistic skills might be strong enough to deploy a wider range of connective devices while B1 learners might simply be more willing to take risks and impress their potential assessors.

Regarding *nonetheless* and *yet*, it is interesting to observe that they follow exactly the same pattern across the three levels. Thus, B2 students refrain from using them while B1 learners employ both with exactly the same low frequency (0.081) and C1 students use it slightly more frequently (0.201) but not particularly so. There are two plausible reasons for this. On the one hand, instructors might refrain from teaching *nonetheless* given that it is also sparsely employed by native and professional writers. For example, in the *British Academic Written English Corpus* (BAWE), *nonetheless* has a frequency of 0.0032 per cent per million words in contrast to *however*, with a frequency of 0.15 per cent. Hence, instructors might consider it too formal and infrequent to be taught. On the

other hand, this group of B2 students might also be lower risk-takers, especially in this specific context of a high-stakes examination (Brown 2001).

The case of *yet* is also worth commenting on. Given the polysemy of the term, which can also be employed as a time adverb, learners might avoid using it as a contrastive adverbial. In contrast, native speakers seem to resort to it so as to widen the range of contrastive devices in their writings (Zhang 2021). In its contrastive use, *yet* is characterised by sentence-initial position, and by being followed by a pause indicated by the punctuation sign of a comma.

Results show that contrastive *yet* is employed just on one occasion by a male B1 learner, as illustrated in example (14). Its ratio is thus extremely low (0.00003):

- (14) Europe has been learnt in the study culture, **yet** how much importance have in our future? [9294M]

In the case of B2, there are no tokens of this connector, while its number increases to three tokens at C1 level, as illustrated in examples (15) and (16), by a female and male learner, respectively:

- (15) **Yet**, a question needs raising here: What kind of city we want? [600004F]
 (16) And **yet**, we vote blind. [900012M]

Albeit at a low frequency, there is nonetheless an increase in the number of tokens, which shows that students at lower and higher levels are either more willing to take risks or more confident in their linguistic ability. At B2 level, they seem to remain more insecure, which could explain why they resort to ‘safer’ options, even if they might be aware of other alternatives.

Against initial expectations, the use of *in contrast* is also extremely low, with a frequency of 0.081 among B1 learners and 0.067 among C1 students. Given the similarity with the Spanish phrase *en contraste*, I expected a positive language transfer from the participants’ L1 to their EFL writings. This might explain why B1 learners employ the phrase even more frequently than C1 participants, as they might be translating it from their Spanish mother tongue. As argued by Wanderley and Demmans (2020), proficiency raises a learner’s awareness of L2 rules and their application while less skilled learners (B1 students in our corpus) tend to employ transfer more frequently. Regarding the use of *still* as a contrastive connector, there are no occurrences in any of the three sub-corpora, which shows that learners may either be unaware of its contrastive meaning or unwilling

to risk its use given that they feel safer employing other alternatives to express contrast, such as *however* or *on the other hand*, in line with previous studies like Zhang (2021), who reports an underuse of *still* and *yet* by his Chinese EFL participants in contrast to native users. Finally, there are no occurrences of *by contrast* in any of the three levels. This absence seems unexpected, as this expression might be more related to the students' own L1. It is difficult to explain why they might be refraining from using it, but it might be due to the belief that this is a calque from Spanish and could hence be negatively assessed by the examiners. In fact, previous research has consistently shown that Spanish EFL learners tend to avoid cognates due to different reasons (e.g., different contextual uses, different frequency of use, different meaning despite a common etymology) to the extent that, according to Scarcella and Zimmerman (2005: 125), "L2 students who speak Spanish as a first language might avoid them in formal English writing" (see also Whitley 2002).

4.2. Gender and use of contrastive linking adverbials

Table 4 summarises the frequency of use according to the participants' gender. To quantify results, all the writings by female and male students have been respectively added up according to the writers' gender, independently of their level. All in all, the female dataset amounts to 29,629 words and the male one to 28,313 words. As in Section 5.1, frequencies have been normalised per thousand words to allow for comparison between the two genders. The column on the left presents the raw number of occurrences while that on the right displays the normalised frequency per gender.

Linking adv.	Female learners		Male learners	
	Freq. (raw)	Freq. (ptw)	Freq. (raw)	Freq. (ptw)
<i>However</i>	78	2.633	43	1.519
<i>Nevertheless</i>	18	0.608	9	0.318
<i>Nonetheless</i>	3	0.101	1	0.035
<i>Yet</i>	1	0.034	2	0.071
<i>In contrast</i>	1	0.034	1	0.035
<i>On the other hand</i>	63	2.126	55	1.943

Table 4: Frequency of use according to gender

As can be observed in Table 4, female learners tend to employ more linking adverbials than their male counterparts, occasionally doubling the latter's use. This seems to align with previous research that claims that female learners tend to employ more connectors

than male learners (Winkler 2008; Tavakoli and Karimnia 2017; Alqahtani and Abdelhalim 2020). A chi-square test, however, has shown that the differences between both genders are not statistically significant ($p=0.419$). In fact, gender difference in discourse markers use has rendered contradictory results. For example, Azeez *et al.* (2023) report a higher use of discourse markers by male than by female participants, although their study is based on the fictional characters of Shaw's theatre play *Arms and the Man* rather than on real speakers/writers. Still, Núñez Mercado (2022), in a more recent study on the written expression of EFL Mexican students, found no significant difference in terms of discourse markers by the two genders, in line with the results found in the current study, where differences are not statistically significant either. Regarding variety, results do not seem affected by the participants' gender, both male and female students resorting mainly to three adverbial linkers, namely, *however*, *on the other hand*, and *nevertheless*, while polysemous expressions like *still* or *yet* are scarcely used or not used at all. Cognates such as *by contrast* or *in contrast* also tend to be avoided by these participants. As already mentioned, this might be due to the students' fear that these expressions are negatively evaluated as calques (Whitley 2002).

4.3 Expressing contrast according to the CEFR/CV

According to the CEFR/CV (2020), the ability to express contrast by means of adverbial links or connectors should already be present at B1 level, the main difference with higher levels being the range of use by learners. In other words, the higher the level, the more varied the connectors are expected to be. More specifically, the CEFR/CV sums up these differences as shown in Figure 1. As already pointed out, the difference between B1 and B2 seems clearer as B1 learners are expected to “use *a limited number* of cohesive devices” while B2 learners should be able to “use *a range* of linking words and cohesive devices” (my emphasis). This difference is vaguer when it comes to the contrast between B2 and C1 students, as the latter are supposed to use “a variety of cohesive devices”, the difference between “a range” and “a variety” arguably unhelpful for learners, evaluators, and teachers.

COHERENCE AND COHESION		PROSIGN
C2	Can create coherent and cohesive text making full and appropriate use of a variety of organisational patterns and a wide range of cohesive devices.	
C1	Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices. Can produce well-organised, coherent text, using a variety of cohesive devices and organisational patterns.	
B2	Can use a variety of linking words efficiently to mark clearly the relationships between ideas. Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse. Though there may be some 'jumpiness' in a long contribution. Can produce text that is generally well-organised and coherent, using a range of linking words and cohesive devices. Can structure longer texts in clear, logical paragraphs.	
B1	Can introduce a counter-argument in a simple discursive text (e.g. with 'however'). Can link a series of shorter, discrete simple elements into a connected, linear sequence of points. Can form longer sentences and link them together using a limited number of cohesive devices, e.g. in a story. Can make simple, logical paragraph breaks in a longer text.	
A2	Can use the most frequently occurring connectors to link simple sentences in order to tell a story or describe something as a simple list of points. Can link groups of words with simple connectors like 'and', 'but' and 'because'.	
A1	Can link words or groups of words with very basic linear connectors like 'and' or 'then'.	
Pre-A1	No descriptors available	

Figure 1: Descriptors for Coherence and Cohesion (CV, 2020: 142)

As already discussed in Section 5.1, *on the other hand* is widely used by B1 and B2 students (2.622 ptw and 2.236 ptw, respectively) in line with previous research such as Rica Peromingo (2012), whose results report the same tendency among these two levels. *However* is also widely employed by the three levels, with a higher frequency by C1 students, also in line with previous research (Zhang 2021). The frequency of more formal contrastive devices such as *nevertheless* and *nonetheless* is reversed when it comes to *nonetheless*, with zero occurrences at B2. Polysemous devices such as *yet* or *still* are only employed by higher level students, although at a low frequency (2%).

The results of the present study do not seem to support this distinction, as B1 learners use as many linking devices as students at B2 and even display a wider variety, akin to that of C1 students. Thus, out of the eight contrastive linking adverbials under scrutiny, B1 and C1 students employ six (*on the other hand*, *however*, *nevertheless*, *nonetheless*, *yet*, and *in contrast*). By contrast, B2 students show the same frequency of their counterparts but a marked lack of variety, with only a repertoire limited to three linking adverbials (*on the other hand*, *however*, and *nevertheless*).

Interestingly, these results seem to point to B2 learners being less ‘adventurous’ than those at lower and higher levels. Indeed, they seem to prefer using safer (but fewer) options which they know well rather than displaying a variety of cohesive devices as the CEFR/CV points out. Unfortunately, and to the best of my knowledge, there is no prior research on the subject. As a result, any conclusions about this group of B2 students remain speculative. Nevertheless, this gap may open up interesting avenues for future research in other text-types (for example, to explore whether this behaviour also appears in genres such as personal emails or narratives, or whether it is specific to more formal argumentative essays). As Tabari (2022) argues, there are other factors that might affect the learners’ writing results. In his study, he found that task planning could mitigate the content and organisation demands on L2 writers’ cognitive processes and consequently enable the allocation of proper attentional resources to other aspects of the writing system and processes, hence enhancing writing products. Since the participants in the present study were under the pressure of taking a high-stakes exam, this might explain why they tend to overuse familiar connectors like *however* even at higher levels of linguistic proficiency like C1.

As already discussed in Section 5.1, *on the other hand* is widely used by B1 and B2 students (2.622 ptw and 2.236 ptw, respectively) in line with previous research such as Rica Peromingo (2012), whose results report the same tendency among these two levels. In a more recent study, Faya-Cerqueiro and Martín Macho-Harrison (2022) find similar results, with *on the other hand* (and wrongly constructed expressions like **in the other hand*) being overwhelmingly employed, not only to show contrast but also to enumerate (often as an adjacent pair with *on the one hand*).

However is also widely employed by the three levels, with a higher frequency by C1 students, also in line with previous research (Zhang 2021). In fact, textbooks may also encourage its use as a contrastive device. For example, the *Oxford EAP B1* textbook (de Chazal and Rogers 2013a: 158) includes the following linking words to express contrast: *but, on the other hand, despite, even if, even though, and however* and the same list is present in the edition for B2 learners (de Chazal and Rogers 2013b). Furthermore, its placement seems rather fixed in initial-sentence position, especially by B1 and B2 learners, rather than expert writers’ rhematic position.

More formal contrastive devices such as *nevertheless* and *nonetheless* are, however, much less frequent and even absent in the case of B2 students. This aligns with previous

research on Spanish EFL writers, who underuse *nevertheless* when compared with native writers. For example, Carrió-Pastor (2013: 196–197) found that her non-native writers employed *nevertheless* in 1.2 per cent of the cases in contrast to native writers (2.3%). Similar results were obtained for *nonetheless*, which although scarcely used by both cohorts, was employed 0.7 per cent by native writers versus 0.4 per cent by non-native ones.

As for polysemous devices such as *yet* or *still*, results show that they are either sparsely used or not used at all by all three groups. This absence might be due to two main reasons. On the one hand, their polysemous nature might confuse learners, who are more familiar with these expressions as time adverbials. On the other, instructors and teaching materials may also play a central role.

The role of instruction should not hence be underestimated, for example, regarding the transferable use of L1 strategies. As argued by Faya-Cerqueiro and Martín Macho-Harrison (2022: n.a), teachers should advocate for the need to adopt a pedagogic approach that favours translanguaging and boosts positive language transfer to “extrapolate certain abilities” already acquired in their L1 to their EFL (see also Cenoz *et al.* 2022). Furthermore, resorting to their knowledge of the L1 might become extremely helpful when tackling writing tasks in line with previous research (see Plata-Ramírez 2016). Fostering students’ metalinguistic reflection might render positive results, especially in the use of cognates like *in contrast* or *by contrast*, which students might be avoiding in the fear that they are calques rather than correct linking devices often employed by native writers.

5. CONCLUSION

The present paper aimed to answer the following research questions, repeated here for the sake of clarity:

RQ1. To what extent (if any) is the use of contrastive linking adverbials affected by the learners’ CEFR level? In other words, to what degree do higher-level students employ a wider range of linking adverbials in contrast to lower-level students? More specifically, I will be analysing two groups of contrastive linking adverbials: (i) the single adverbs *however*, *nonetheless*, *nevertheless*, *yet* and *still*; and (ii) the prepositional phrases *on the other hand*, *by contrast*, and *in contrast*.

RQ2. To what extent (if any) does the learners' gender affect their use of these contrastive linking adverbials (i.e., frequency of use and syntactic variety of use)?

In response to the first question, results show that there is a statistically significant relationship between the students' level and their use of contrastive linking adverbials, which would confirm the first hypothesis. Nevertheless, there is a progressive decrease in terms of frequency in the use of linking adverbials as students' level increases. Thus, normalised frequencies show that B1 learners use a further amount of linking adverbials (4.753 ptw), in contrast to B2 (4.311 ptw) and C1 (3.491 ptw). This could be due to the fact that higher-level students resort to a wider variety of linguistic strategies to express contrast, including both coordination and subordination, which future research intends to explore. Regarding variety, and against expectations, students at B1 level display the same range of linking adverbials than those at C1 (albeit at lower frequencies), while B2 students' repertoire is limited to three of the eight expressions under analysis. Polysemous devices such as *still* and *yet* seem too complex for students to employ, also independently of the level. This could be due to the fact that the learners of the present dataset were taking a high-stakes exam, which might lead them to more conservative and 'safer' options rather than more 'adventurous' ones. As already mentioned, the role of instruction should not be underestimated either.

Regarding the second question, gender does not seem to play any role in the frequency and range of these learners' use of contrastive devices. In fact, despite minor differences which are not statistically significant, both male and female learners seem to resort to the same kind of devices at the three levels, even if female writers tend to a higher frequency of use.

Admittedly, this paper is not without limitations, especially as regards the size of the dataset and the fact that I have focused exclusively on eight linking adverbials and not on other linguistic devices to express contrast, such as subordinating conjunctions like *although*, or coordinating ones like *but*, which have been shown to be extensively used at B1 level (see Lahuerta Martínez 2002). On the other hand, the present dataset has not been contrasted with an L1 corpus (e.g., the BAWE) to observe whether native writers also underuse or overuse these linking adverbials. Future research from a cross-linguistic approach might shed light in this regard.

REFERENCES

- Ahmed, Abdelhamid M., Xiao Zhang, Lameya M. Rezk and William S. Pearson. 2023. Transition markers in Qatari university students' argumentative writing: A cross-linguistic analysis of L1 Arabic and L2 English. *Ampersand* 10: 100110.
- Alqahtani, Sahar Nafel and Safaa M. Abdelhalim. 2020. Gender-based study of interactive metadiscourse markers in EFL academic writing. *Theory and Practice in Language Studies* 10/10: 1315–1325.
- Altenberg, Bengt and Marie Tapper. 1998. The use of adverbial connectors in advanced Swedish learners' written English. In Sylviane Granger ed. *Learner English on Computer*. London: Longman, 80–93.
- Appel, Randy. 2020. An exploratory analysis of linking adverbials in post-secondary texts from L1 Arabic, Chinese, and English writers. *Ampersand* 7: 100070.
- Azeez, Abbas Ibrahim, Ayad Hameed Mahmoud and Ahmed Adel Nouri. 2023. A multi-perspective study of discourse markers: An attempt to sort out the muddle among EFL teachers-students. *EDUCASIA: Jurnal Pendidikan, Pengajaran, dan Pembelajaran* 8/1: 25–48.
- Azkarai, Agurtzane. 2015. L1 use in EFL task-based interaction: A matter of gender? *European Journal of Applied Linguistics* 3/2: 159–179.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 2021. *Grammar of Spoken and Written English*. Amsterdam: John Benjamins.
- Blakemore, Diane. 2006. Discourse markers. In Laurence R. Horn and Gregory Ward eds. *The Handbook of Pragmatics*. London: Blackwell Publishing, 221–240.
- Brown, H. Douglas. 2001. *Teaching by Principles: An Interactive Approach to Language Pedagogy*. New York: Addition Wesley: Longman, Inc.
- Carrió-Pastor, María Luisa. 2013. A contrastive study of the variation of sentence connectors in academic English. *Journal of English for Academic Purposes* 12/3: 192–202.
- Celce-Murcia, Marianne and Diane Larsen-Freeman. 1999. *The Grammar Book*. Boston: Heinle & Heinle Publishers.
- Cenoz, Jasone, Oihana Leonet and Durk Gorter. 2022. Developing cognate awareness through pedagogical translanguaging. *International Journal of Bilingual Education and Bilingualism* 25/8: 2759–2773.
- Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
- Council of Europe. 2020. *Companion Volume*. Cambridge: Cambridge University Press.
- Cowan, Ron. 2008. Discourse connectors and discourse markers. In Ron Cowan ed. *The Teacher's Grammar of English*. Cambridge: Cambridge University Press, 615–634.
- de Chazal, Edward and Louis Rogers. 2013a. *Oxford EAP. Intermediate B1+*. Oxford: Oxford University Press.
- de Chazal, Edward and Louis Rogers. 2013b. *Oxford EAP. Upper Intermediate B2*. Oxford: Oxford University Press.
- Díez Prados, Mercedes. 2010. Gender and L1 influence on EFL learners' lexicon. In Rosa M^a Jiménez Catalán ed. *Gender Perspectives on Vocabulary in Foreign and Second Languages*. London: Palgrave Macmillan, 44–73.
- Faya-Cerqueiro, Fátima and Ana Martín Macho-Harrison. 2022. Uso de conectores en la redacción de textos argumentativos: Comparación entre L1 y L2. *Ocnos. Journal of Reading Research* 21/2.

- Ghadessy, Mohsen, Robert L. Roseberry and Alex Henry. 2008. *Small Corpus Studies and ELT*. Amsterdam: John Benjamins.
- Granger, Sylviane and Stephanie Tyson. 1996. Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes* 15/1: 17–27.
- Granger, Sylviane. 2015. Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1/1: 7–24.
- Gilquin, Gaëtanelle and Magali Paquot. 2008. Too chatty: Learner academic writing and register variation. *English Text Construction* 1/1: 41–61.
- Ha, Myung-Jeong. 2016. Linking adverbials in first-year Korean university EFL learners' writing: A corpus-informed analysis. *Computer Assisted Language Learning* 29/6: 1090–1101.
- Hasselgren, Angela. 1994. Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics* 4/2: 237–258.
- Hosseinpour, Rasoul Mohammad and Hossein Hosseini Pour. 2022. Adversative connectors use in EFL and native students' writing: A contrastive analysis. *The Electronic Journal for English as a Second Language* 26/1.
- Huddleston, Rodney and Geoffrey Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
- Jiménez Catalán, Rosa María and Julieta Ojeda Alba. 2014. Diagnóstico de las dificultades en el uso de los conectores en una tarea escrita por parte de aprendices de inglés como lengua extranjera: Estrategias de corrección. *Didáctica [Lengua y Literatura]* 26: 197–217.
- Lahuerta Martínez, Ana Cristina. 2002. The use of discourse markers in EFL learners' writing. *Revista Alicantina de Estudios Ingleses* 15/4: 123–132.
- Lee, Kent. 2013. Korean ESL learners' use of connectors in English academic writing. *English Language Teaching* 25/2: 81–103.
- Lee, Kent. 2020. Chinese ESL writers' use of English contrastive markers. *English Language Teaching* 32/4: 89–110.
- Leedham, Maria and Guozhi Cai. 2013. *Besides ... on the other hand*: Using a corpus approach to explore the influence of teaching materials on Chinese students' use of linking adverbials. *Journal of Second Language Writing* 22/4: 374–389.
- Modhish, Abdulhafeed Saif. 2012. Use of discourse markers in the composition writings of Arab EFL learners. *English Language Teaching* 5/5: 56–61.
- Mora Díaz, Luz Mary and Yeimmy Gómez Orjuela. 2021. Understanding the English language through a creative writing workshop: Adjectives and adverbs essential for EFL learners. *Shimmering Words: Research and Pedagogy E-journal* 11: 52–73.
- Narita, Masumi, Chieko Sato and Masatoshi Sugiura. 2004. Connector usage in the English essay writing of Japanese EFL learners. *Language Resources and Evaluation Conference* 27/1: 1171–1174.
- Navarro Gil, Noelia and Helena Roquet Pugès. 2020. Linking or delinking of ideas? *Revista Española de Lingüística Aplicada* 33/2: 505–535.
- Neff-van Aertselaer, JoAnne and Emma Dafouz-Milne. 2008. Argumentation patterns in different languages: An analysis of metadiscourse markers in English and Spanish texts. In Martin Putz and JoAnne Neff eds. *Developing Contrastive Pragmatics Interlanguage and Cross-cultural Perspectives*. Berlin: Mouton de Gruyter, 87–102.
- Núñez Mercado, Carlos. 2022. Male and female BA students' use of discourse markers: A corpus-based study. *Lenguas en Contexto* 13: 4–12.

- Park, Yong-Yae. 2013. Korean college EFL students' use of contrastive conjunctions in argumentative writing. *English Teaching* 68(2): 263–284.
- Pérez-García, Elisa and M^a Jesús Sánchez. 2020. Emotions as a linguistic category: Perception and expression of emotions by Spanish EFL students. *Language, Culture and Curriculum* 33/3: 274–289.
- Pérez-Paredes, Pascual, María Sánchez-Tornel and José M^a Alcaraz Calero. 2012. Learners' search patterns during corpus-based focus-on-form activities: A study on hands-on concordancing. *International Journal of Corpus Linguistics* 17/4: 482–515.
- Plata-Ramírez, José Miguel. 2016. Language switching: Exploring writers' perceptions on the use of their L1s in the L2 writing process. *Revista Internacional de Lenguas Extranjeras* 5: 47–77.
- Porras Wadley, Luis. 2022. *The Ultimate CAE Writing Guide for C1 Cambridge*. Granada: KSE Academy.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
- Rica Peromingo, Juan Pedro. 2012. Corpus analysis and phraseology. *Linguistics and the Human Sciences* 6/3: 321–343.
- Scarcella, Robin Cameron and Cheryl Boyd Zimmerman. 2005. Cognates, cognition, and writing: An investigation of the use of cognates by university second-language learners. In Andrea Tyler, Mari Takada, Yiyoun Kim and Diana Marinova eds. *Language in Use: Cognitive and Discourse Perspectives on Language and Language Learning*. Washington, D.C.: Georgetown University Press, 123–136.
- Schenck, Andrew. 2020. Examining the influence of native and non-native English-speaking teachers on Korean EFL writing. *Asian-Pacific Journal of Second and Foreign Language Education* 5/1: 2–17.
- Tabari, Mahmoud Abdi. 2022. Investigating the interactions between L2 writing processes and products under different task planning time conditions. *Journal of Second Language Writing* 55: 100871.
- Tankó, Gyula. 2008. Composition: The use of adverbial connectors in Hungarian university students' argumentative essays. In John Sinclair ed. *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins, 157–181.
- Tavakoli, Mahboobeh and Amin Karimnia. 2017. Dominant and gender-specific tendencies in the use of discourse markers: Insights from EFL learners. *World Journal of English Language* 7/2: 1–9.
- Wanderley, Leticia Farias and Carrie Demmans Epp. 2020. Identifying negative language transfer in writing to increase English as a Second Language learners' metalinguistic awareness. In Vitomir Kovanović, Maren Scheffel, Niels Pinkwart and Katrien Verbert eds. *Companion Proceedings of the 10th International Conference on Learning Analytics and Knowledge (LAK20)*. Frankfurt: The University of Frankfurt, 722–725.
- Whitley, Melvin Stanley. 2002. *Spanish/English Contrasts: A Course in Spanish Linguistics*. Washington, D.C.: Georgetown University Press.
- Winkler, Elizabeth Grace. 2008. A gender-based analysis of discourse markers in Limonese Creole. *Sargasso: A Journal of Caribbean Literature, Language & Culture*, Special Issue: *Linguistic Explorations of Gender & Sexuality* 1: 53–72.
- Yilmaz, Ercan and Kenan Dikilitas. 2017. EFL learners' uses of adverbs in argumentative essays. *Novitas-ROYAL (Research on Youth and Language)* 11/1: 69–87.

- Yoon, Choongil. 2019. On the other hand: A comparative study of its use by Korean EFL students, NS students and published writers. *Journal of Language Sciences* 26/2: 273–297.
- Zhang, Yan. 2021. *Adversative and Concessive Conjunctions in EFL Writing: Corpus-based Description and Rhetorical Structure Analysis*. Berlin: Springer.

Corresponding author

Carmen Maíz-Arévalo
Complutense University of Madrid
Department of English Studies
Plaza Menéndez Pelayo s/n
ES-28040 Madrid
Spain
E-mail: cmaizare@ucm.es

received: February 2025
accepted: May 2025

A corpus-based study on the transitive uses of English physiological verbs

Beatriz Rodríguez Arrizabalaga
University of Huelva / Spain

Abstract – This paper examines the transitivity potential of a group of English unergative verbs that denote physiological processes, a syntactico-semantic verbal class which has not received enough attention in the literature. Through a qualitative and quantitative corpus-based analysis of 24 verbs conducted on the *Corpus of Contemporary American English* (COCA), *British National Corpus* (BNC) and *Corpus of Global Web-Based English* (GloWbE), it will be shown that the syntactic flexibility of this verbal class is higher than stated in previous studies since, in addition to the cognate object construction (*Burp the same garlic burps*), the substance object construction (*Breathe the smoke, Lucien*), and the resultative construction (*He yawned open his mouth*), these verbs have been documented in seven other transitive patterns in which they increase their valency with the addition of a non-canonical direct object: *x's way* constructions (*I sweated my way through a painful run*), reaction object constructions (*Emma hiccups a yes*), caused-motion constructions (*She spits phlegm into a Kleenex*), the preposition drop object alternation (*He shit the rug*), the understood body-part object alternation (*The elk snuffled her face through the snow*), away constructions (*Everyone laughs the evening away*), and causative constructions (*Let's burp this baby!*).

Keywords – physiological verbs; constructions; transitivity alternations; non-canonical objects; corpus study

1. INTRODUCTION

Transitivity is a complex linguistic phenomenon involving several syntactic, semantic, and pragmatic aspects, which has always occupied a central place in the studies of any language in the world. Indeed, as Dixon (1979: 102), Næss (2007: 1–2) and La Polla *et al.* 2011: 469), among others, remark, transitivity is a universal phenomenon, widely investigated over time from multiple theoretical perspectives, which, broadly speaking, has been approached from two main different perspectives.

In the studies that follow the traditional definition of the term (Bloomfield 1933; Chomsky 1965; Stockwell *et al.* 1973; Comrie 1981; Radford 1988), transitivity is considered an either-or phenomenon which differentiates sharply between transitive and intransitive verbs, depending only on their complementation by a direct object. According

to Radford (1988: 42), verbs “can be classified as transitive or intransitive depending on whether or not they require a direct object in order to form a grammatically complete sentence.”

On the contrary, in later studies (Hopper and Thompson 1980; Givón 1984; Tsunoda 1985; Davidse 1991; Taylor 1995; Hale and Keyser 2002; Næss 2007), transitivity is understood as a multifaceted, gradual phenomenon determined by several factors, apart from the presence of an object, which affects the whole clause:

Transitivity involves a number of components, only one of which is the presence of an object of the verb. These components are all concerned with the effectiveness with which an action takes place, e.g., the punctuality and telicity of the verb, the conscious activity of the agent, and the referentiality and degree of affectedness of the object. (Hopper and Thompson 1980: 251)

Because of this change in the study of transitivity, at the end of the twentieth century, from the 1980s onwards, several studies have been carried out to examine the different alternations in which the same verb takes part and, consequently, on its possible transitive and intransitive uses (Hale and Keyser 1986; Croft 1991; Devís Márquez 1993; Levin 1993; Amberber 1996; Martínez Vázquez 1998).

As far as English is concerned, Levin’s (1993) study, which develops around the hypothesis that the meaning of a verb determines its syntactic behavior, is particularly interesting as it provides a fairly complete syntactico-semantic classification, not only of English verbs, but also of the alternations they enter. Nevertheless, 32 years after its publication, Levin’s (1993) work is in need of revision since, as the author foresaw at the time (1993: 17), there are verb classes and alternations that require further investigation to which, I believe, corpus studies can make a great contribution:

Many of the diathesis alternations and verb classes are familiar and well studied. Others have received relatively little attention, and I hope that their inclusion may stimulate further study. (Levin 1993: 17)

A good case in point are English verbs that denote physiological processes which, despite describing basic bodily functions essential to the life of living beings, have not been fully addressed in the literature (Thalberg 1972; Snell-Hornby 1983; Fellbaum 1990; McClure 1990; Dixon 1991; van Gelderen 2018). These are the focus of this paper, whose main objective is to study, through a qualitative and quantitative corpus-based analysis, the alternations with direct object complementation which this semantic class of English

unergative verbs enters.¹ The specific research questions that this study seeks to answer are, thus, the following:

RQ1. Which are the transitive structures in which English verbs denoting physiological processes participate?

RQ2. Do physiological verbs behave uniformly in them?

The paper is structured as follows. Section 2 provides a review of English physiological verbs. Section 3 explains the methodology underlying my analysis, with special emphasis on the rationale behind the selection of the verbs studied and the corpora used—the *Corpus of Contemporary American English* (COCA), the *British National Corpus* (BNC94) and the *Corpus of Global Web-Based English* (GloWbE)—, as well as on the search parameters for data extraction and interpretation. Section 4 presents and discusses the most significant findings. Finally, Section 5 offers some concluding remarks.

2. PREVIOUS STUDIES ON ENGLISH PHYSIOLOGICAL VERBS

The verbs that express physiological processes (*cough*, *sweat*, *sneeze*, etc.) constitute an interesting area of study since they are unique in describing basic and essential bodily functions. However, research on this type of English verbs has been overlooked in the literature, being the subject of study in only a small number of works (Thalberg 1972; Snell-Hornby 1983; Fellbaum 1990; McClure 1990; Dixon 1991; van Gelderen 2018).

Semantic studies highlight the heterogeneity of this verbal class, as it comprises verbs which refer to actions that are both unconsciously (e.g. *blushing*) and consciously (e.g. *spitting*) controlled. However, since most of the actions they denote occur without the subject's willingness, their inherent involuntariness is considered their distinguishing feature (Thalberg 1972: 57). For Snell-Hornby (1983) and Dixon (1991), who use the labels 'descriptive' and 'corporeal' verbs, respectively,² these verbs fall into a category

¹ As postulated in the 'Unaccusative Hypothesis', originally formulated within Relational Grammar by Perlmutter (1978) and later reviewed by Burzio (1986) in Government and Binding Theory, intransitive verbs do not show a uniform syntactic, semantic and aspectual behavior, hence the distinction established between unaccusative and unergative intransitive verbs. For their grammatical behavior in English, see, among others, Levin and Rappaport Hovav (1995), Kuno and Takami (2004), and Kijparnich (2011).

² See also in this regard Kudrnáčová (2005), which deals with the group of 'corporeal' verbs that denote nonvolitional oscillatory movements of different body parts.

where the focus is on the internal causation of the actions described, which means that their subject is at the same time the doer and the experiencer of the processes they convey.

As far as their aspectual properties are concerned, physiological verbs also show mixed behavior. As Comrie (1976) remarks, they describe atelic actions (e.g. *sweating*) that spread over an indefinite time period, as well as telic actions (e.g. *sneezing*) which, in contrast, express short and punctual actions.

As regards their syntactic characterization, most studies emphasize their intransitive character (example (1)), though acknowledging their possible transitive use in colloquial or figurative expressions (example (2)), and in causative constructions (example (3)).

- (1) Jake **hiccuped** softly.³ (COCA: Fiction)
- (2) Dad **coughed up green phlegm**. (COCA: Web)
- (3) Stacy **had burped the baby**. (COCA: Fiction)

Physiological verbs have been likewise analyzed from a pragmatic dimension given their socio-cultural connotations. Since certain bodily functions (e.g. *burping*) are considered impolite or taboo, the use of the verbs that express them is restricted by social and cultural norms, as well as by context (Snell-Hornby 1983; Levinson 1983; Wierzbicka 1997; Leech *et al.* 2001; Allan and Burridge 2006). Therefore, according to Lakoff's (1973) politeness theory, their use is allowed without any restrictions in informal settings, whereas in formal environments the employment of euphemisms (e.g. *expel air*) is preferred.

Levin's (1993) work stands out among all the studies mentioned above because, as far as I know, it is the most comprehensive syntactico-semantic study on English physiological verbs. The author labels them as 'Verbs of Bodily Processes' and includes them with seven other verbal classes in the larger group of 'Verbs Involving the Body.'⁴ According to Levin (1993: 217–219), the verbs denoting bodily processes can be further subdivided into three different classes, attending to their specific syntactic behavior:

³ For reasons of space and clarity, most of the corpus examples provided have been adapted and are presented in a shortened version.

⁴ The seven remaining verbal classes are Verbs of Nonverbal Expression, Verbs of Gesture/Signs Involving Body Parts, 'Snooze' Verbs, 'Flinch' Verbs, Verbs of Body-Internal States of Existence, 'Suffocate' Verbs, and Verbs of Bodily State and Damage to the Body. For the specific verbs in each group, see Levin (1993: 219–227).

- (i) ‘Hiccup’ verbs (*belch, blush, burp, hiccup, pant, sneeze, sniffle, snore, snuffle, swallow, wheeze, yawn*), which relate to involuntary bodily processes that are not (typically) under the control of the person that experiences them.
- (ii) ‘Breathe’ verbs (*bleed, breathe, cough, cry, dribble, drool, puke, spit, sweat, vomit, ?weep*). Except for *breathe*, which “also describe[s] taking air into the body,” all these verbs “relate to emitting a substance from the body” (Levin 1993: 218).
- (iii) ‘Exhale’ verbs (*exhale, inhale, perspire*), which, despite being semantically related to ‘breathe’ verbs, show a limited set of properties, owing to their Latinate origin.

As shown in Table 1, for Levin (1993: 217–219), these three verbal classes do not function alike in the cognate object construction, the resultative construction, and the substance object construction, and neither do they all take *at* and *on* complements, or have zero-related nominals:

	‘Hiccup’ verbs	‘Breathe’ verbs	‘Exhale’ verbs
Zero-related nominals	<i>A hiccup.</i>	<i>A breath.</i>	<i>*An exhale.</i>
At complementation	<i>*Paul hiccuped at Mary.</i>	<i>*Paul breathed at Mary.</i>	<i>*Paul exhaled at Mary.</i>
On complementation	<i>*Paul hiccuped on Mary.</i>	<i>Paul breathed on Mary.</i>	<i>*Paul exhaled on Mary.</i>
Cognate object construction	<i>?Paul hiccuped a loud hiccup.</i>	<i>Paul breathed a deep breath. vs. *Paul sweated a cold sweat.</i>	<i>*Paul exhaled a deep breath.</i>
Resultative construction	<i>??Paul hiccuped himself sick.</i>	<i>*Paul breathed Mary awake.</i>	
Substance object construction		<i>The dragon breathed fire.</i>	

Table 1: Verbs of bodily processes (Levin 1993: 217–219)

3. METHODOLOGY

Transitive uses of English unergative verbs of physiological processes are studied in the present research through a qualitative and quantitative, corpus-based analysis of 24

different verbs describing bodily processes in American and British English. These are the two ‘native’ varieties of English most commonly used worldwide and with the greatest influence on the different World Englishes (Hickey 2012: 1; Han 2019: 93).⁵

The corpora chosen for the analysis are the *Corpus of Contemporary American English* (COCA; Davies 2008) and the *British National Corpus* (BNC; Davies 2007), two balanced corpora which include a wide range of text categories representative of the linguistic varieties under consideration. Since the BNC does not include Internet records, I have resorted to the British section of the *Corpus of Global Web-Based English* (GloWbE; Davies 2013), which contains more than 387 million words, to cover this particular register in the British variety of English. Despite the semantic problems Levin’s (1993) study entails, for the selection of the verbs analyzed I have taken this work as my starting point, given that in it these verbs conform a specific class.⁶ Of the 26 verbs included in Levin’s (1993) work (see Section 2 above), I have excluded *swallow* due to its transitive nature, and the verbs in the partially synonym pairs *belch/burp*, *pant/wheeze*, *sniffle/snuffle*, *cry/weep*, *dribble/drool*, and *puke/vomit* with the highest relative frequency in the corpora, because of the elevated number of tokens analyzed (91,964 tokens in total);⁷ the only exception is the verb *perspire*, which is included despite being synonymous with *sweat*, as well as *exhale* and *inhale*, due to their Romance origin.

In addition to these, I have also analyzed five other unergative verbs which, due to their meaning, should be included within the class of ‘breathe’ verbs. On the one hand, *laugh*, for being the antonym of *weep*; on the other hand, four verbs referring to basic physiological processes (*pee*, *urinate*, *shit*, and *excrete*) which, quite surprisingly, are absent from Levin’s (1993) work. The Latinate origin of *urinate* and *excrete*, however, leads me to classify them as ‘exhale’ verbs. The full list of verbs, together with their raw and relative frequencies, is provided in Table 2.

⁵ Information provided by *WorldData* (2025).

⁶ A case in point is the pair of verbs *cough* and *burp*, which could be well included within the same group. However, in Levin’s (1993) work, without any reason provided, they are classified, respectively, as a ‘breathe’ and a ‘hiccup’ verb.

⁷ The raw and relative frequencies of the verbs discarded are indicated between brackets: *belch* (COCA: 938/0.94; BNC: 142/1.42; GloWbE: 225/0.58), *pant* (COCA: 4,781/4.81; BNC: 351/3.51; GloWbE: 383/0.99), *sniffle* (COCA: 1,229/1.24; BNC: 21/0.21; GloWbE: 71/0.18), *cry* (COCA: 68,156/68.63; BNC: 5,535/55.35; GloWbE: 17,465/45.06), *dribble* (COCA: 2,038/2.05; BNC: 187/1.87; GloWbE: 1,238/3.19), and *vomit* (COCA: 4,725/4.76; BNC: 417/4.17; GloWbE: 1,786/4.61).

Verbal class		Raw frequency			Relative frequency		
		COCA	BNC	GloWbE	COCA	BNC	GloWbE
'Hiccup' verbs	<i>Blush</i>	3,804	781	1,050	3.83	7.81	2.71
	<i>Burp</i>	981	40	176	0.99	0.40	0.45
	<i>Hiccup</i>	369	18	45	0.37	0.18	0.12
	<i>Sneeze</i>	2,443	150	578	2.46	1.50	1.49
	<i>Snore</i>	3,239	259	742	3.26	2.59	1.91
	<i>Snuffle</i>	277	59	85	0.28	0.59	0.22
	<i>Wheeze</i>	1,421	130	229	1.43	1.30	0.59
	<i>Yawn</i>	2,763	395	713	2.78	3.95	1.84
	<i>Bleed</i>	14,881	797	4,036	14.99	7.97	10.41
	<i>Breathe</i>	39,303	2,929	8,372	39.58	29.29	21.60
'Breathe' verbs	<i>Cough</i>	9,127	866	2,022	9.19	8.66	5.22
	<i>Drool</i>	1,837	78	552	1.85	0.78	1.42
	<i>Laugh</i>	111,455	8,785	22,931	112.24	87.85	59.16
	<i>Pee</i>	6,101	125	1,045	6.14	1.25	2.70
	<i>Puke</i>	2,095	43	310	2.11	0.43	0.80
	<i>Shit</i>	2,368	69	530	2.38	0.69	1.37
	<i>Spit</i>	12,845	962	2,997	12.94	9.62	7.73
	<i>Sweat</i>	9,950	641	1,893	10.02	6.41	4.88
	<i>Weep</i>	8,021	1,073	2,307	8.08	10.73	5.95
	<i>Excrete</i>	667	122	222	0.67	1.22	0.57
'Exhale' verbs	<i>Exhale</i>	3,296	96	237	3.32	0.96	0.61
	<i>Inhale</i>	5,954	295	1,071	6.00	2.95	2.76
	<i>Perspire</i>	417	45	66	0.42	0.45	0.17
	<i>Urinate</i>	1,529	89	664	1.59	0.89	1.71

Table 2: List of English unergative verbs of bodily processes analyzed

To exclude as many intransitive uses as possible, I have resorted to the proximity criterion of the corpora, which allows the combination of the lemma (the verbs chosen) with a particular word class. Due to the nominal nature of direct objects, on the one hand, and to the frequent modification of cognate objects, on the other, the noun category was selected over other word categories listed in the corpora, placing it in an interval of three spaces to the right of the verbs.⁸ This facilitated the retrieval of objects with pre- and post-modification, such as the ones illustrated in (4) and (5).

(4) He **sneezed a light sneeze**. (COCA: Academic)

(5) They **wept tears of laughter**. (BNC: Fiction)

This search parameter has, furthermore, yielded hits of resultative and caused-motion constructions, as in examples (6)–(7), which, despite having a pronominal object, are interesting for this study owing to their transitive nature.

(6) He **yawned himself to sleep**. (COCA: Fiction)

⁸ It should be noted that the cognate object construction is one of the transitive structures compatible with English unergative verbs (Massam 1990; Dixon 1991; Levin and Rappaport Hovav 1995; Macfarland 1995; Mittwoch 1998; Felser and Wanner 2001; Kuno and Takami 2004; Kim and Lim 2012; van Gelderen 2018).

- (7) You'd like the judge **to laugh you into jail!** (COCA: Web)

The tokens retrieved were manually analyzed so as to discard those examples with complex verbs, such as *snuffle up* in (8), and those with a postverbal noun phrase displaying a syntactic function different from that of direct object, like the ones in examples (9)–(11), where they function, respectively, as subjects, subject complements, and adverbials.

- (8) I **snuffled up** a deep breath of it. (COCA: Fiction)
- (9) Below me **yawns a raw mineral landscape**. (COCA: Magazine)
- (10) She **blushed a sudden agonized red**. (COCA: Web)
- (11) He **snored all night**. (COCA: Newspaper)

Once these examples had been rejected, the valid matches were analyzed in their context to exclude, on the one hand, those where the verb is used metaphorically, as in the example with an inanimate subject illustrated in (12),⁹ and, on the other, those instances which, despite manifesting a literal meaning, are not related to any kind of physiological process; two cases in point are examples (13) and (14), where *cough* behaves as a verb of nonverbal expression, and *inhale* expresses the action of 'eating'.

- (12) **The generator hiccupped** some bad diesel. (COCA: Fiction)
- (13) [...] the professor **coughing the right answer**. (COCA: Spoken)
- (14) I just **inhaled peanutbutter pie for breakfast**. (COCA: Blog)

The results of this manual process are summarized in Table 3.

⁹ For the metaphorical extensions of the verbs denoting bodily processes, see Lakoff and Johnson (1980), where it is shown how the body serves as a source domain for the conceptualizing of abstract, non-physical actions such as the release of excessive information (*He vomited the facts all at once; She coughed up the answer*).

Verbal class		Examples analyzed (v – n)	Transitive patterns attested
‘Hiccup’ verbs	<i>Blush</i>	3,833	11 (0.28%)
	<i>Burp</i>	539	70 (12.98%)
	<i>Hiccup</i>	142	11 (7.74%)
	<i>Sneeze</i>	930	34 (3.65%)
	<i>Snore</i>	1,223	15 (1.22%)
	<i>Snuffle</i>	186	23 (12.36%)
	<i>Wheeze</i>	738	31 (4.20%)
	<i>Yawn</i>	1,087	28 (2.57%)
	<i>Bleed</i>	7,426	135 (1.81%)
	<i>Breathe</i>	16,941	4,345 (25.64%)
‘Breathe’ verbs	<i>Cough</i>	4,080	167 (4.09%)
	<i>Drool</i>	1,055	43 (4.07%)
	<i>Laugh</i>	21,995	867 (3.94%)
	<i>Pee</i>	2,906	302 (10.39%)
	<i>Puke</i>	878	37 (4.21%)
	<i>Shit</i>	1,183	390 (32.96%)
	<i>Spit</i>	7,719	1,184 (15.33%)
	<i>Sweat</i>	5,353	239 (4.46%)
	<i>Weep</i>	3,779	275 (7.27%)
	<i>Excrete</i>	742	321 (43.26%)
‘Exhale’ verbs	<i>Exhale</i>	3,516	482 (13.70%)
	<i>Inhale</i>	3,982	1,994 (50.07%)
	<i>Perspire</i>	180	2 (1.11%)
	<i>Urinate</i>	1,005	32 (3.18%)
TOTAL		91,964	11,038 (12.00%)

Table 3: Number of examples analyzed and transitive patterns attested

4. FINDINGS AND DISCUSSION

The first conclusion derived from my analysis is that the grammatical behavior of the English verbs denoting bodily processes is not so heterogeneous as postulated by Levin (1993), since most of them have been attested in the cognate object, the resultative, and the substance object constructions, as illustrated in Table 4.

Verbal class		Cognate object constructions	Resultative constructions	Substance object constructions
'Hiccup' verbs	<i>Blush</i>	–	1	–
	<i>Burp</i>	1	2	13
	<i>Hiccup</i>	2	1	2
	<i>Sneeze</i>	5	2	4
	<i>Snore</i>	3	–	2
	<i>Snuffle</i>	1	–	10
	<i>Wheeze</i>	7	1	3
	<i>Yawn</i>	9	5	–
	<i>Bleed</i>	55	38	22
'Breathe' verbs	<i>Breathe</i>	1,607	3	2,556
	<i>Cough</i>	15	12	81
	<i>Drool</i>	9	2	21
	<i>Laugh</i>	506	168	7
	<i>Pee</i>	13	2	63
	<i>Puke</i>	–	3	24
	<i>Shit</i>	4	4	124
	<i>Spit</i>	3	1	580
	<i>Sweat</i>	31	10	119
'Exhale' verbs	<i>Weep</i>	213	14	22
	<i>Excrete</i>	2	–	307
	<i>Exhale</i>	137	–	304
	<i>Inhale</i>	107	–	1,853
	<i>Perspire</i>	–	–	2
	<i>Urinate</i>	1	–	29
TOTAL		2,731 (24.52%)	269 (2.41%)	6,148 (55.20%)

Table 4: Raw frequencies of physiological verbs in cognate object, resultative and substance object constructions in COCA, BNC, and GloWbE

In addition to these, seven other transitive patterns, absent from Levin's (1993) description, have been also attested, which reveals that these verbs are much more flexible syntactically than originally expected. In three of them (the *x's way*, the caused-motion, and the reaction object constructions), the three verbal classes analyzed have likewise considerable presence (see Table 5). This outcome is not surprising because these structures are, as resultatives, prototypical of satellite-frame languages, such as English, owing to the incorporation of a manner component into the verbs of their primary predications (Talmy 1985: 2000).

Verbal class		<i>X</i> ' way constructions	Reaction object constructions	Caused-motion constructions
'Hiccup' verbs	<i>Blush</i>	1	5	—
	<i>Burp</i>	4	2	3
	<i>Hiccup</i>	1	5	—
	<i>Sneeze</i>	5	—	16
	<i>Snore</i>	1	—	—
	<i>Snuffle</i>	7	1	—
	<i>Wheeze</i>	16	3	1
	<i>Yawn</i>	10	—	—
	<i>Bleed</i>	2	5	2
	<i>Breathe</i>	8	43	123
'Breathe' verbs	<i>Cough</i>	15	11	30
	<i>Drool</i>	1	5	3
	<i>Laugh</i>	76	6	56
	<i>Pee</i>	—	2	1
	<i>Puke</i>	2	2	6
	<i>Shit</i>	—	—	4
	<i>Spit</i>	4	34	559
	<i>Sweat</i>	33	2	11
	<i>Weep</i>	1	10	4
	<i>Excrete</i>	—	1	11
'Exhale' verbs	<i>Exhale</i>	—	26	13
	<i>Inhale</i>	1	—	27
	<i>Perspire</i>	—	—	—
	<i>Urinate</i>	—	—	—
TOTAL		188 (1.68%)	163 (1.49%)	870 (7.81%)

Table 5: Raw frequencies of physiological verbs in *x*'s way, reaction object and caused-motion constructions in COCA, BNC, and GloWbE

In the four remaining patterns (the preposition drop object alternation, 'away' constructions, the understood body-part object alternation, and causative constructions), however, their occurrence is more restricted, as shown in Table 6.

Verbal class		Preposition drop object alternation	Understood body-part object alternation	<i>Away</i> constructions	Causative constructions
'Hiccup' verbs	<i>Blush</i>	1	–	–	3
	<i>Burp</i>	1	–	–	44
	<i>Hiccup</i>	–	–	–	–
	<i>Sneeze</i>	1	–	1	–
	<i>Snore</i>	1	–	8	–
	<i>Snuffle</i>	–	4	–	–
	<i>Wheeze</i>	–	–	–	–
	<i>Yawn</i>	–	2	2	–
	<i>Bleed</i>	–	–	8	3
	<i>Breathe</i>	–	1	4	–
'Breathe' verbs	<i>Cough</i>	–	–	3	–
	<i>Drool</i>	–	–	2	–
	<i>Laugh</i>	10	1	37	–
	<i>Pee</i>	219	–	2	–
	<i>Puke</i>	–	–	–	–
	<i>Shit</i>	254	–	–	–
	<i>Spit</i>	2	–	1	–
	<i>Sweat</i>	16	–	6	11
	<i>Weep</i>	7	–	4	–
	<i>Excrete</i>	–	–	–	–
'Exhale' verbs	<i>Exhale</i>	2	–	–	–
	<i>Inhale</i>	6	–	–	–
	<i>Perspire</i>	–	–	–	–
	<i>Urinate</i>	2	–	–	–
TOTAL		522 (4.68%)	8 (0.07%)	78 (0.70%)	61 (0.54%)

Table 6: Raw frequencies of physiological verbs in the preposition drop object alternation, the understood body-part object alternation, and *away* and causative constructions in COCA, BNC, and GloWbE

The presence of English physiological verbs in the ten transitive structures illustrated in Table 7 confirm, in the end, de Swart's (2007: 16) hypothesis about the possible transitivity of intransitive verbs, thus suggesting, in another respect, that transitivity may be a potential feature of any verb, as stated in Roberge (2002) and Bilous (2012).

Patterns	Number of verbs attested	Numbers of examples attested
Substance object construction	22	6,148 (55.20%)
Cognate object constructions	21	2,731 (24.52%)
<i>X's way</i> constructions	18	188 (1.68%)
Caused-motion constructions	17	870 (7.81%)
Resultative constructions	17	269 (2.41%)
Reaction object constructions	17	163 (1.49%)
Preposition drop object alternation	13	522 (4.68%)
<i>Away</i> constructions	12	78 (0.70%)
Causative constructions	4	61 (0.54%)
Understood body-part object alternation	4	8 (0.07%)

Table 7: Number of verbs and examples found in the transitive patterns attested

Therefore, due to their clear binary syntactic status, I consider it more appropriate to classify English physiological verbs, following Bouso (2021), as ‘amphibious’ rather than as unergative verbs.¹⁰

4.1. Cognate object constructions

Despite coming from the internal accusative structure characteristic of classical languages, cognate object constructions have also been studied in relation to various modern languages, such as English (Bassols de Climent 1945; Rodríguez Adrados 1992; Bary and de Swart 2005). For this reason, I follow the broad definition of English cognate objects, initially provided by Sweet (1891), which is still pertinent nowadays (Kuno and Takami 2004; de Swart 2007; Sailer 2010; Wilson 2019). According to Sweet,

Sometimes an intransitive verb is followed by a noun in the common form which repeats the meaning of the verb, as in *sleep the sleep of the just*, *fight a good fight*, where the noun is simply the verb converted into a noun, and in *fight a battle*, *run a race*, where the noun repeats the meaning, but not the form, of the verb. Such object-nouns are called cognate objects. (Sweet 1891: 91)

Therefore, I have considered as such not only those noun phrases morphologically related to the verb of the sentence, as Levin (1993) does, but also those that are semantically connected with it.¹¹

Almost all the verbs examined have been found complemented by morphological cognate objects. The only exceptions are the ‘hiccup’ verbs *blush*, *hiccup*, and *snuffle*, the ‘breathe’ verb *puke*, and the ‘exhale’ verbs *excrete*, *inhale*, and *perspire*. In some cases, morphological cognates are separated from the verbs by a punctuation mark, either by a comma, as in examples (15)–(16) or by an en dash, as in examples (17)–(18), intended, in my view, to soften the redundancy they entail as morphological repetitions of the verb, along with the usual, but not obligatory, modification they adopt (Massam

¹⁰ The term ‘amphibious verbs’, which I borrow from Visser (1963-1973), is alternatively referred to in the literature as ‘bivalent verbs’ (Rivas 1996), ‘ambitransitive verbs’ (Dixon and Aikhenvald 2000), ‘dual transitivity verbs’ (Huddleston and Pullum 2002), and ‘labile verbs’ (McMillion 2006).

¹¹ In studies which defend that cognates must be morphologically related to the verb, the objects which are semantic repetitions of the intransitive verbs that they complement constitute a different class of objects, variously named in the literature as ‘transitivizing objects’ (Massam 1990), ‘hyponyms of Cos’ (Dixon 1991; Felser and Wanner 2001), ‘hyponymous’ or ‘hyponymic objects’ (Hale and Keyser 2002; Real Puigdollers 2008), or simply ‘non-Cos’ (Ogata 2011). Though differentiated, Wilson (2019) classifies both kinds as ‘Inclusive Objects’ due to the similarities they share.

1990; Dixon 1991; Levin 1993; Macfarland 1995; Mittwoch 1998; Felser and Wanner 2001; Nakajima 2006; Höche 2009).

- (15) Capona **wheezed, a breathless, flat wheeze**. (COCA: Fiction)
- (16) He started **coughing, a cough that was to be** persistent. (GloWbE: Web)
- (17) He heard them **breathing –one breath long and light**. (BNC: Fiction)
- (18) She **was weeping now –a painful mute weeping**. (COCA: Fiction)

Moreover, two out of the seven aforementioned verbs have been attested with semantic cognate objects, namely, *breath* as complement of *hiccup* and *inhale*, and *feces* as the complement of *excrete*, as illustrated in examples (19)–(21).

- (19) She **hiccupped a breath**. (COCA: Fiction)
- (20) She **inhales a shuddery breath**. (COCA: Fiction)
- (21) The human body **excretes feces**. (COCA: Web)

In addition to these three verbs, semantic cognates have also been documented with several of the verbs analyzed that accept morphological cognate complementation. The antonym verbs *laugh* and *weep* are especially interesting in this regard as they have been found with more than one semantic cognate; *laugh*, specifically, with *guffaw*, *grin*, and *giggle* ((22)–(24)), and *weep*, with *tears* and *sobs* (examples (25)–(26)):

- (22) She **laughs, a big-bellied guffaw**. (COCA: Fiction)
- (23) Demon tryed **not to laugh himself a sight evil grin** on his face. (GloWbE: Web)
- (24) He will then **laugh his false giggle**. (GloWbE: Blog)
- (25) She had **wept more tears** over the loss of dear ones [...]. (COCA: Web)
- (26) She **wept choked, snotty sobs** [...]. (COCA: Fiction)

Furthermore, my analysis has brought about some examples, unnoticed in the literature, which I call ‘understood cognate object constructions’, given that they have a cognate object semantically implicit, but not overtly expressed. Two different classes are to be distinguished: those built around the ‘breathe’ verbs *weep*, *sweat*, and *pee*, in which the cognate object, if present, is encoded as a nominal postmodifier headed by *of* (see (27)–

(33)), and those constructed around the ‘exhale’ verb *inhale*, that have, in turn, temporal objects which are originally the postmodifiers of the missing cognate object (*air*), omitted together with the preposition that heads them (*of*) (see (34) and (35)).

- (27) She **wept a river (of tears)**, poor woman. (COCA: Movies)
- (28) And **sweat rivers (of sweat)**. (COCA: Magazine)
- (29) He **peed a river (of pee)** on the floor (COCA: Blog)
- (30) I **wept oceans (of tears)**. (COCA: Web)
- (31) I **sweat a swimming pool (of sweat)** onto the floor. (COCA: Blog)
- (32) A small dog **peed a bright yellow puddle (of pee) at the base of the linden tree**. (COCA: Fiction)
- (33) I **sweat buckets (of sweat)** when I run. (GloWbE: Web)
- (34) You **can inhale (the air of) summer**. (COCA: Magazines)
- (35) She **inhales (the air of) the deep blue morning**. (COCA: Fiction)

The examples in the former class (examples (27)–(33)) additionally have an idiomatic meaning of excessive sorrow, as they all are variations of the phraseological expression *cry a river*, thus calling into question the rigid fixation of idiomatic expressions (Fernando and Flavell 1981). Notice that in them the verb *cry* is replaced by other ‘breathe’ verbs that denote the emission of a liquid substance (*weep*, *sweat*, and *pee*), and the object *river* is replaced with other liquid containers (*oceans*, *swimming pools*, *puddle*, *buckets*) which, depending on their size, express the intensity with which the verbal action has been carried out, thus behaving as quantity adverbials.

4.2. Resultative constructions

Resultatives are a prototypical pattern of Germanic, satellite-framed languages, in which two different predications are merged in one simple sentence; a primary verbal predication that describes how the change of state denoted in the secondary predication (of adjectival or prepositional nature) is achieved. Surprisingly, however, the corpora provided no examples of the group of ‘exhale’ verbs, which come from Latin, another satellite-framed language (Talmy 2000: 104).

The classes of ‘hiccup’ and ‘breathe’ verbs, on the contrary, do frequently appear in resultatives, having yielded 269 examples in which only *snore* and *snuffle* have not been attested. The resultative examples attested are especially revealing for two reasons. Firstly, they show that English prepositional resultatives, which have received almost no attention in the literature until recent times (Beavers 2002; Riaubiené 2015; Flach 2020), are as common as adjectival ones, which, in opposition, have been deeply studied for being considered the prototypical ones (Peña Cervel 2009: 758; Riaubiené 2015: 65). To my knowledge, only those introduced by prepositions denoting a goal like *to* and *into* (Beavers 2002: 17) and *till* and *until* (Riaubiené 2015: 73–74) have been recently investigated for being the most common ones. In my analysis, in fact, all the prepositional resultatives attested are headed by the prepositions *to* and *into*; to the following illustrated cases (*to bits* (example (36)), *into snot* (example (37)), *to death* (example (38)), and *into a coma* (example (39))), many others have to be added (*into the deepest, most profound state of hypnosis, into a swoon, into convulsions, to helplessness, into stitches, into a fit of coughing, into such a roaring mirth, to health, to tears, and into exhaustion*). Moreover, most of them have been found as complements of the verb *laugh*.

(36) You **could burp it to bits**. (COCA: Movies)

(37) She **hiccupped her tears into snot**. (COCA: Blog)

(38) You **didn’t drool yourselves to death**. (COCA: Movies)

(39) I **am trying not to laugh myself into a coma**. (COCA: Blog)

Secondly, they demonstrate that adjectival resultatives are not so restricted as Goldberg (1995: 195) states.¹² In addition to those shown in examples (40)–(42), some other adjectival resultatives have been found: *delirious* has been attested in conjunction with *cough*; *dry* with *weep*; *clean*, *slick*, and *silly* with *sweat*; and *weak*, *hoarse*, *senseless*, *breathless*, *limp*, *slim*, *purple*, *incontinent*, *stupid*, *sore*, and *wet* with *laugh*. They all are, however, non-gradable adjectives with a clear delimited lower bound, thus satisfying the necessary condition to enter the English resultative construction.

(40) My intestines would have heard that dirty old song and dance enough **to puke a mountain full**. (COCA: Blog)

¹² For Goldberg (1995: 195), the list of possible adjectival resultatives is limited to *asleep/awake*, *open/shut*, *flat/straight/smooth*, *free/full/empty*, *dead/alive*, *sick*, *hoarse*, and *crazy*.

(41) He's **gon na sweat himself unsterile**. (COCA: TV)

(42) My grandam **wept herself blind** at my parting. (COCA: Movies)

4.3. Substance object constructions

The findings concerning the substance object construction are especially remarkable since they reveal that not only 'breathe' verbs ((43)–(44)), as Levin (1993: 218) remarks, but also the 'hiccup', ((45)–(46)), and 'exhale' ((47)–(48)) classes occur in this transitive pattern. Except for the 'hiccup' verbs *blush* and *yawn*, all the verbs analyzed have been documented in this structure complemented by a wide range of substances (6,148 instances).

(43) He **drooled saliva and blood and something like water**. (COCA: Fiction)

(44) I **puked tequila** in the parking lot. (COCA: TV)

(45) And **sneeze cocaine** all over the room. (COCA: Magazine)

(46) He's **not snoring fire**. (COCA: TV)

(47) You've **excreted your food and drink**. (COCA: Magazine)

(48) [...] by making you **urinate more liquid than you drink**. (COCA: Magazine)

Some instances of the substance object construction with the synonymous verbs *breathe* and *exhale* are worth noticing since, instead of an agent, as in the previous ones, they have a locative adverbial promoted to subject position, as shown in (49a) and (50a). In them, this agentive participant (with generic reference, as seen in (49b) and (50b)), is semantically implicit, but syntactically unexpressed. Consequently, they could be considered a special kind of Levin's (1993: 82) 'Location Subject Alternation,' related in her study just to the group of 'Fit' verbs, illustrated in (51a) and (51b):

(49a) **The whole district exhales the hospitality of a grave**. (COCA: Fiction)

(49b) **One/You exhale(s) the hospitality of a grave in the whole district**.

(50a) **Sao Paulo breathes street art on every corner**. (GloWbE: Web)

(50b) **One/You breathe(s) street art on every corner in Sao Paulo**.

(51a) **We sleep five people in each room**. (Levin 1993: 82)

(51b) **Each room sleeps five people.**4.4. *X's way constructions*

In its canonical form [SUB_i [V [POSS_i *way*] OBL]], the *x's way* construction also merges two different predicative relationships in one simple sentence: a primary predication which basically describes how the movement denoted in the secondary predication is performed.¹³ Considered to be another diagnosis of unaccusativity in English (Jackendoff 1990; Marantz 1992; Levin 1993; Levin and Rappaport Hovav 1995; Ausensi 2019), the attestation of unergative verbs of bodily processes in this construction is not surprising. In fact, all the verbs examined, except for the 'breathe' verbs *pee* and *shit*, and the 'exhale' verbs *excrete*, *exhale*, *perspire*, and *urinate*, have been documented in it (188 cases).

The examples of the *x's way* construction attested in my analysis bring to light several aspects of interest for the study of this construction: firstly, that the noun *way* allows adjectival modification, (example (52)). In agreement with Jackendoff (1990: 217) and Goldberg (1995: 206), but in opposition to McColm (2019: 245), this observation demonstrates that the noun *way* is the head of a referential object, and not just a meaningless syntactic marker of the construction; secondly, it also shows that its final directional can be realized not only in the form of prepositional phrases, as in (53), but also in that of adverb phrases, as in (54); and finally, it provides evidence that the movement denoted, which does not necessarily have an endpoint as seen in the examples attested with the preposition *through* ((55)) (Hilpert 2014: 38), can be physical (either spatial, as in (56), or temporal, as in (57)), as well as metaphorical (as in (58)).

(52) Entwhistle will enjoy his publicly-funded retirement by **laughing all his masonic way to the bank**. (GloWbE: Web)

(53) He **coughed his way to an early grave**. (COCA: Fiction)

(54) The old guard **will be yawning his way out** while the new guard **yawns his way in**. (COCA: Fiction)

(55) Americans sniffle and **sneeze their way through a billion colds every year**. (COCA: Spoken)

(56) She was **blushing her way back to their table**. (COCA: Fiction)

¹³ For the different meanings of the *x's way* construction, see Goldberg (1995), Israel (1996), and Perek (2018).

(57) I **wheezed my way through long blue-black nights**. (COCA: Fiction)

(58) I want to **burp my way to victory**. (COCA: Movies)

4.5. Reaction object constructions

The reaction object construction describes how the reaction of its subject participant, “an emotion or disposition” in Levin’s (1993: 98) words, is expressed. Therefore, the verbs that enter this construction, a member of the classes of nonverbal of expression verbs ((59)), manner of speaking verbs ((60)), and verbs of gestures and signs ((61)) (Levin 1993: 98; Huddleston and Pullum 2002: 305), adopt an extended meaning which can be paraphrased as ‘expressing a reaction by V-ing’ (Levin 1993: 98).

(59) Pauline **smiled her thanks**. (Levin 1993: 98)

(60) She **mumbled her adoration**. (Levin 1993: 98)

(61) I **nodded my agreement**. (Huddleston and Pullum 2002: 305)

The scope of this structure should be extended, however, to physiological verbs since all the verbs analyzed, except for the ‘hiccup’ verbs *sneeze*, *snore*, and *yawn*, the ‘breathe’ verb *shit*, and the ‘exhale’ verbs *inhale*, *perspire*, and *urinate* have been found in the reaction object construction (163 cases).¹⁴ The examples retrieved verify, furthermore, the description of canonical reaction objects provided by Martínez Vázquez (2014: 186–188). On the one hand, they are either directly linked to the verb without any determiner mediating between them ((62)), or introduced by the indefinite article ((63)), or a possessive determiner coreferential with the clausal subject ((64)).

(62) That cat **pukes pure hatred**. (COCA: Web)

(63) She **blushes a ‘thank you’**. (COCA: Fiction)

(64) They all **laughed their relief**. (BNC: Fiction)

¹⁴ A similar position is adopted by Bouso (2021: 292), who notes that “the R[eaction]O[bject]C[onstruction] has kept attracting more and more verb types.” In addition to verbs of gestures (21 types), Bouso (2021) analyzes five other different verbal classes in American English: verbs of sound emission and related verbal classes (38 types), verbs of bodily processes (5 types), verbs of instrument of communication (7 types), verbs of activity (6 types), and verbs of light emission (3 types). Of the verbs of bodily processes examined in the present paper, only *bleed*, *breathe*, and *snuffle* are included in Bouso’s (2021: 295) analysis.

On the other, they consist of nominalized conventional speech-act formulae, like *goodbyes* (example (65)), nouns derived from expressive illocutionary verbs, such as *complaint* (example (66)), or attitudinal nouns which, by disclosing an emotional state of the mind, like *panic* (example (67)), lead Martínez Vázquez (2014: 176) to name this pattern the ‘expressive object construction’.

(65) One by one they **wept their goodbyes**. (COCA: Fiction)

(66) He **wheezed a complaint**. (COCA: Fiction)

(67) His hazel eyes **bleed panic**. (COCA: Fiction)

However, the corpus data reveal that the range of reaction objects that complement physiological verbs is greater than originally expected, since they far outnumber those reported by Levin (1993: 98),¹⁵ as also confirmed in Martínez Vázquez’s (2014) study. In addition to the reaction objects previously exemplified, some others have been attested: *appreciation, amusement, lament, sorrow, support, emotion, fear, and frustration*, among others.

4.6. *Caused-motion constructions*

As Goldberg (1995: 152–153) states, the caused-motion construction expresses, in its canonical form [SUBJ [V OBJ OBL]], the movement of its object participant along a path, as a consequence of the particular manner in which the clausal subject performs the verbal action. Due to its syntactico-semantic status, this construction has been deeply investigated in the literature on resultatives (Goldberg 1991, 1995; Boas 2003; Goldberg and Jackendoff 2004; Peña Cervel 2009). For Goldberg (1995: 81–89), for instance, resultatives are a metaphorical extension of the caused-motion pattern on the basis that result phrases are figurative types of goals. Therefore, as stated by Peña Cervel (2009: 758), who establishes a cognitive continuum between both constructions, “the difference comes to the fore if we analyze the resultant state of these two configurations.” In other words, a change of location in caused-motion expressions, and a change of state in resultatives.

¹⁵ In Levin’s (1993: 98) work, the possible reaction objects in English are restricted to *approval, disapproval, assent, admiration, disgust, yes, and no*.

Similarly to resultatives and to the *x's way* and reaction object constructions,¹⁶ the caused-motion pattern is also prototypical of the satellite-framed linguistic typology in which two different events are chained to each other. As expected, therefore, the verbs studied are also relatively frequently documented in this pattern (870 instances). In fact, only the ‘hiccup’ verbs *blush*, *hiccup*, *snore*, *snuffle*, and *yawn*, and the ‘exhale verbs’ *perspire* and *urinate* have not been attested in my analysis in the caused-motion construction.

The examples documented show, furthermore, that the path denoted by the final prepositional phrase in this pattern, usually headed by the prepositions *across* ((68)), *out of* ((69)), *onto* ((70)), and most commonly *into* ((71)), specifies the source ((69)) or the goal ((70)–(71)) of the caused-motion event, though sometimes both can be overtly expressed, as in example (72).

- (68) I **burped that meatball right across the room**. (COCA: Movies)
- (69) They’d **laugh me straight out of the door**. (GloWbE: Web)
- (70) They **spit the husks onto the visiting officers’ uniforms**. (GloWbE: Web)
- (71) Men **excrete their bodily fluids into women**. (COCA: *Blog*)
- (72) [...] as Yogi **sneezing a large grub from his nostril right into the audience**. (GloWbE: Web)

4.7. The preposition drop object alternation

Some of the verbs studied have also been found complemented by a direct object of circumstantial nature, either locative or temporal, which, by losing the head of the prepositional phrase functioning as adverbial from which it derives, ceases to be a peripheral argument to display a central function in the sentence (522 cases). They fit, thus, into Levin’s (1993: 43–44) ‘Preposition Drop Alternations’ and Esquivel Rodríguez’s (2010: 162) ‘Constructions with promoted direct object’ (*Construcciones con promoción a objeto directo*).

This finding is very significant since none of the two types of preposition drop alternations identified by Levin (1993: 43–44) is related to physiological verbs. The

¹⁶ According to Peña Cervel (2009: 743), the *x's way* construction is, in fact, “a more specific and constructionally conventionalized version of the more generic caused-motion configuration.”

‘Locative Preposition Drop Alternation’ is found with certain motion verbs that take directional phrases as complements, as in (73),¹⁷ and the ‘With Preposition Drop Alternation,’ in turn, with the small set of reciprocal ‘meet’ verbs that entail some kind of social interaction, as in (74).

(73) Martha **climbed (up) the mountain**. (Levin 1993: 43)

(74) Jill **met (with) Sarah**. (Levin 1993: 44)

The locative direct objects attested, where items of clothing are included, are completely affected by the bodily process described in the sentence. They receive, thus, the same holistic interpretation that Levin (1993: 43) attributes to the objects denoting paths or goals that complement motion verbs, and which contrasts with their partitive meaning when functioning as adverbials. As seen in examples (75)–(78), this kind of promoted object has been documented with verbs of scatological processes (*pee*, *shit*, *sweat*, and *urinate*), and the ‘breathe’ verbs *bleed*, *spit*, and *weep*.

(75) I’d **rather shit (in) my underwear!** (COCA: Web)

(76) It is possible on match days to see people **urinating (in) the streets**.
(GloWbE: Web)

(77) I **bleed (in) every fucking place**. (COCA: Movies)

(78) If you’re guilty of murdering Kaplan, even **spitting (on) the sidewalk**, [...].
(COCA: Movies)

In addition to these examples, which clearly fit into Levin’s (1993: 43) ‘Locative Preposition Drop Alternation’, the verbs examined have yielded several tokens of the pattern which, following Levin’s (1993: 43) terminology, I label ‘Temporal Preposition Drop Alternation’. Notice that *during* is usually the preposition dropped in them. It has, however, a more restricted use than its locative variant, having only been attested with the ‘hiccup’ verb *blush*, and the ‘breathe’ verbs *laugh*, *sweat*, and *weep*. As examples (79)–(82) show, it usually has eventive objects (*interview*, *episode*, *conversation*)

¹⁷ Specifically, the classes of *run* verbs, as shown in (73), verbs of vehicle names (*They skated (along) the canals*), and some verbs of inherently directed motion (*Matha slowly descended (down) the stairs*; see Levin 1993: 43).

premodified by the adjective *entire*, which suggests that they all are also totally affected by the physiological process described in the clause.¹⁸

(79) I **blushed the entire interview**. (COCA: Magazine)

(80) I **laughed the entire episode!!!** (GloWbE: Blog)

(81) I was **sweating the entire conversation**. (GloWbE: Web)

(82) In the rubble, Mohammed Afzl **wept (all) Tuesday** for his brother. (COCA: News)

The holistic/partitive distinction that differentiates the two patterns entering the locative preposition drop alternation is, thus, valid as well to account for its temporal counterpart.

4.8. The understood body-part object alternation

According to Levin (1993: 34–35), the understood body-part object alternation is only compatible with the verbs describing conventionalized gestures and signs called ‘wink’ verbs ((83)), and the verbs of body care belonging to the group of ‘floss’ verbs ((84)). As seen in these two examples, its object, which refers to the body-part directly involved in the verbal action described in the sentence, is semantically understood, and not overtly expressed in its intransitive use.

(83) The departing passenger **waved (his hand) at the crowd**. (Levin 1993: 34)

(84) I **flossed (my teeth)**. (Levin 1993: 34)

Curiously enough, though denoting bodily processes, the verbal class analyzed has not been frequently attested in this transitive pattern. Only the ‘hiccup’ verbs *hiccup*, *snuffle*, and *yawn*, and the ‘breathe’ verbs *breathe* and *laugh* have been documented in the eight examples of the understood body-part object retrieved. Three of them are provided in examples (85)–(87).

(85) **Yawning its oversized jaws** to show him its tusks. (COCA: Fiction)

¹⁸ Besides this interpretation (*I blushed/laughed/was sweating while the interview/episode/conversation was taking place*), which is the most logical one in my view, there is another one, suggested by one of the reviewers, in which the event denoted by the object triggers the corporeal process (*The entire interview/episode/conversation made me blush/laugh/sweat*). Though context should be key to decode the meaning of these structures, it does not resolve the ambiguity in these examples.

- (86) Whether you **breathe your nose or mouth**, to find the approach that works best for you. (GloWbE: Web)
- (87) We see childless women on TV, **laughing bright red lips** as they stare through glass. (COCA: Fiction)

The examples with the verb *puke* in (88) and (89) show, nevertheless, that not all the sentences with a body-part complementing their verbs can be considered representatives of this particular alternation, as the body-part in them is not directly involved in the physiological process described. What they do, in my view, is to convey some nuances of excess and exaggeration that point out the great extent to which the verbal action has been carried out, thus behaving similarly to quantity adverbials.

- (88) I'd sooner **puke my intestines** than see you naked! (COCA: TV)
- (89) I'm gon na **puke my guts**. (COCA: Movies)

4.9. 'Away' constructions

According to Jackendoff (1997), the 'time'-*away* construction is characterized by the presence of a volitional subject that uses the time encoded in its direct object by performing the verbal action described in the sentence. Since only intransitive verbs can appear in this alternation, the unergative verbs of bodily processes studied are very suitable candidates to participate in such pattern. Nevertheless, only the 'hiccup' verbs *sneeze*, *snore*, and *yawn*, and the 'breathe' verbs *bleed*, *cough*, *drool*, *laugh*, *spit*, and *sweat* have yielded examples of this construction. As shown below, the particle *away* can precede or follow its temporal object; usually a part of the day ((90)), a season ((91)), the noun *life* ((92)), or any other temporal expression ((93)).

- (90) Everyone **laughs the evening away** and enjoys themselves. (COCA: Web)
- (91) The classic cartoon image of a big bear **snoring away the winter** in a huge cave [...]. (COCA: Magazine)
- (92) I won't have her **sweating her life away** in the potato fields. (COCA: Fiction)
- (93) He must be dotting around on a cane, **drooling the tiny days away**. (COCA: Fiction)

The restricted presence of physiological verbs in the 'time'-*away* construction is a rather surprising fact. This structure, which shares many properties with resultatives and *x'way*

constructions, where these verbs are quite recurrent, is “a distinct member of a family of constructions to which all three belong” (Jackendoff 1997: 534).

The corpus data have yielded some other examples with the ‘breathe’ verbs *bleed*, *breathe*, *drool*, *laugh*, *sweat*, and *weep* that, contrary to Jackendoff (1997) and in agreement with Kim (2010), I also consider special instances of the ‘time’-*away* construction. This is so because their direct objects do not refer to any time expression, but to negative feelings, like *pain* ((94)), *frustration* ((95)), *anger* ((96)), and *guilt* ((97)), called ‘difficulties’ by Kim (2010: 131). Hence, my proposal to consider them representatives of the pattern which I name the ‘feeling’-*away* construction.

(94) You could **bleed the pain away**. (COCA: Movies)

(95) I needed **to breathe away my frustration**. (GloWbE: Web)

(96) You’d **laugh your anger away**. (COCA: Web)

(97) He is free to **weep away his guilt** on her breasts. (COCA: Fiction)

Consequently, its interpretation is richer than that of the ‘time’-*away* construction, since the time consumed performing the verbal action, implicit in its meaning, is spent getting rid of the negative feelings conveyed by its direct object. Therefore, along the line of Kim (2010: 132), who observes that “the ‘time’-*away* construction should extend the range of meaning into the time and the difficulties,” I consider that both patterns are subclasses of a more general structure, which I call the ‘*away*-construction’.

4.10. Causative constructions

In the literature on causative constructions, two different transitive alternations are distinguished (Comrie 1976, 1985; Haspelmath 1993; Levin 1993; Song 1996; Payne 1997). On the one hand, the ‘causative/inchoative alternation’, roughly associated to verbs of change of state and position, as illustrated in (98a–98b); on the other hand, the ‘induced action alternation’, entered, in turn, only by a subset of the ‘run’ verbs, as shown in (99a–99b).

(98a) Janet **broke the cup**. (Levin: 1993: 29)

(98b) **The cup broke**.

(99a) Sylvia **jumped the horse** over the fence. (Levin: 1993: 31)

(99b) **The horse jumped** over the fence.

Though the verbs studied do not seem to have any place in them, 61 causative examples with the ‘hiccup’ verbs *blush* and *burp*, and the ‘breathe’ verbs *bleed* and *sweat* have been attested in my analysis. A small sample of them is provided in examples (100)–(103). This result deserves special attention because it gives clear evidence that some physiological verbs can enter this pattern, if as members of Levin’s (1993: 31–32) catchall category ‘Other instances of causative alternations,’ which comprises a wide range of intransitive verbs.¹⁹ Except for those of the ‘suffocate’ class, these verbs describe internally controlled actions which in certain circumstances can be externally caused and controlled, thus behaving as transitive verbs.

(100) Paul and Poly **were blushing red cheeks** in front of the grandparents.
(COCA: Fiction)

(101) Always **burp your baby** when feeding time is over. (COCA: Web)

(102) The crab can scrape and **bleed my hands** if I do not wear gloves. (COCA: Fiction)

(103) not for the purpose of using his extra knowledge and skill **to sweat his fellow-workman**. (GloWbE: Web)

Additionally, the examples retrieved confirm Levin’s (1993: 32) hypothesis about the semantic restrictions that operate on the objects of this causative pattern. In fact, the verb *blush* has only been found complemented by objects referring to the body part *cheeks*; the objects that complement *burp* allude just to animate beings of a small or young age (*baby, newborn, child, puppies*);²⁰ the verb *bleed* has only been attested with the noun *wound* or the limbs *hands* and *legs*; and finally, any kind of human beings have been documented as objects of *sweat*.²¹

¹⁹ The intransitive verbs included in this category are some verbs of sound, light and substance emission, some spatial configuration verbs, some ‘lodge’ and ‘suffocate’ verbs, and others where *bleed* and *burp* are comprised.

²⁰ My results broaden the scope of Smith’s (1970) findings, as she restricts the object of the causative structures with *burp* exclusively to babies.

²¹ It should be noticed that the object of the causative verb *sweat* can also be inanimate; generally, ingredients in recipes: *Sweat the courgettes in a pan with some olive oil and sliced garlic* (GloWbE: Web), *And sweat the onions until they start to turn translucent (about 6 minutes)* (GloWbE: Web).

5. CONCLUDING REMARKS

The present research has provided a qualitative and quantitative, corpus-based analysis of the potential transitive uses of three different classes of English unergative verbs denoting physiological processes, a verbal class, which, despite describing essential and basic bodily functions, have not been fully addressed in the literature due to its taboo nature. The three verbal classes studied comprise the ‘hiccup’ verbs *blush*, *burp*, *hiccup*, *sneeze*, *snore*, *snuffle*, *wheeze*, and *yawn*, the ‘breathe’ verbs *bleed*, *breathe*, *cough*, *drool*, *laugh*, *pee*, *puke*, *shit*, *spit*, *sweat*, and *weep*, and the ‘exhale’ verbs *excrete*, *exhale*, *inhale*, *perspire*, and *urinate*. Specifically, 91,964 tokens extracted from the COCA, BNC and the British section of the GloWbE corpus have been manually analyzed so as to exclude, on the one hand, as many intransitive examples as possible, and on the other, those instances with complex verbs and those with metaphorical meanings.

After the manual analysis of the tokens retrieved, the evidence presented has revealed that the syntactic flexibility of this English verbal group in terms of the number of transitive patterns in which its members may occur in is higher than has been stated in previous studies, as their presence has been confirmed not only in the cognate object (2,731 examples; 24.52%), the resultative (269 instances; 2.41%), and the substance object constructions (6,148 cases; 55.20%), as Levin (1993: 217–219) indicates, but also in seven other different transitive structures in which they increase their valency with the addition of a non-subcategorized direct object (see Table 7 in Section 4 above): namely, *x’s way* constructions (188 examples; 1.68%), reaction object constructions (163 instances; 1.49%), and caused-motion constructions (870 cases; 7.81%), where they are relatively frequently attested, and the preposition drop object (522 examples; 4.68%) and the understood body-part object (eight cases; 0.07%) alternations, where their occurrence is, as well as in the *away* (78 examples; 0.70%) and the causative constructions (61 cases; 0.54%), more restricted. Consequently, owing to the clear binary syntactic nature of this semantic class of English verbs, I propose their classification as ‘amphibious’ rather than as unergative verbs.

REFERENCES

- Allan, Keith and Kate Burridge. 2006. *Forbidden Words: Taboo and the Censoring of Language*. Cambridge: Cambridge University Press.

- Amberber, Mengistu. 1996. *Transitivity Alternations, Event-types and Light Verbs*. Quebec, Canada: The McGill University dissertation.
- Ausensi, Josep. 2019. Revisiting the elasticity of verb meaning and the way-construction in English. In M. Teresa Espinal, Elena Castroviejo, Manuel Leonetti, Louise McNally and Cristina Real-Puigdollers eds. *Proceedings of Sinn und Bedeutung* 23/1: 75–92.
- Bary, Corien and Peter de Swart. 2005. Additional accusatives in Latin and Ancient Greek: Arguments against arguments. In Judit Gervain ed. *Proceedings of the Ninth ESSLLI Student Session*. Edinburgh: online, 12–24. <https://hdl.handle.net/2066/40222> (15 May, 2024.)
- Bassols de Climent, Mariano. 1945. *Sintaxis Histórica de la Lengua Latina*. Barcelona: CSIC.
- Beavers, John. 2002. Aspect and the distribution of prepositional resultative phrases in English. *LinGO Working Paper* #2002–7. Stanford: Stanford University. https://www.researchgate.net/profile/John-Beavers/publication/2893783_Aspect_and_the_Distribution_of_Prepositional_Resultative/links/0deec522f3c542eb86000000/Aspect-and-the-Distribution-of-Prepositional-Resultative.pdf (20 April, 2024.)
- Bilous, Rostylav. 2012. Transitivity revisited: An overview of recent research and possible solutions. *Proceedings of the 2012 Annual Conference of the Canadian Linguistic Association*: 1–14. https://cla-acl.ca/pdfs/actes-2012/Bilous_2012.pdf (10 April, 2024.)
- Bloomfield, Leonard. 1933. *Language*. New York: Holt, Rinehart and Winston.
- Boas, Hans C. 2003. *A Constructional Approach to Resultatives*. Stanford: CSLI Publications.
- Bouso, Tamara. 2021. *Changes in Argument Structure. The Transitivity Reaction Object Construction*. Bern: Peter Lang.
- Burzio, Luigi. 1986. *Italian Syntax: A Government-binding Approach*. Dordrecht: Kluwer.
- Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, Mass.: The MIT Press.
- Comrie, Bernard. 1976. The syntax of causative constructions: Cross-language similarities and divergences. In Masayoshi Shibatani ed. *The Grammar of Causative Constructions*. New York: Academic Press, 261–312.
- Comrie, Bernard. 1981. *Language Universals and Linguistic Typology*. Chicago: The University of Chicago Press.
- Comrie, Bernard. 1985. Causative verb formation and other verb-deriving morphology. In Tim Shopen ed. *Language Typology and Syntactic Description 3. Grammatical Categories and the Lexicon*. Cambridge: Cambridge University Press, 309–348.
- Croft, William. 1991. *Syntactic Categories and Grammatical Relations: The Role of Semantic Typology*. Chicago: The University of Chicago Press.
- Davidse, Kristin. 1991. Transitivity/Ergativity: The Janus-headed grammar of actions and events. In Martin Davies and Louise Ravelli eds. *Advances in Systemic Linguistics*. London: Pinter, 105–135.
- Davies, Mark. 2008. *Corpus of Contemporary American English*. <https://www.english-corpora.org/coca/> (February–July, 2024.)
- Davies, Mark. 2007. *British National Corpus*. <https://www.english-corpora.org/bnc/> (February–July, 2024.)
- Davies, Mark. 2013. *Corpus of Global Web-Based English*. <https://www.english-corpora.org/glewbe/> (February–July, 2024.)

- de Swart, Petrus Jacobus Franciscus. 2007. *Cross-Linguistic Variation in Object Marking*. Utrecht: LOT.
- Devís Márquez, P. Pablo. 1993. *Esquemas Sintáctico-semánticos: El Problema de las Diátesis en Español*. Cádiz: Servicio de Publicaciones de la Universidad de Cádiz.
- Dixon, Robert M. W. 1979. Ergativity. *Language* 55/1: 59–138.
- Dixon, Robert M. W. 1991. *A New Approach to English Grammar, on Semantic Principles*. Oxford: Oxford University Press.
- Dixon, Robert M. W. and Alexandra Y. Aikhenvald eds. 2000. *Changing Valency: Case Studies in Transitivity*. Cambridge: Cambridge University Press.
- Esquivel Rodríguez, Leo. 2010. Operaciones de aumento de valencia sintáctica en español. *Letras* 48: 151–167.
- Fellbaum, Christiane. 1990. English verbs as a semantic net. *International Journal of Lexicography* 3: 278–301.
- Felser, Claudia and Anja Wanner. 2001. The syntax of cognate and other unselected objects. In Nicole Dehé and Anja Wanner eds. *Structural Aspects of Semantically Complex Verbs*. Bern: Peter Lang, 105–130.
- Fernando, Chitra and Roger Flavell. 1981. *On Idiom: Critical Views and Perspectives*. Exeter: University of Exeter.
- Flach, Susanne. 2020. The emergence of the into-causative: Constructionalization and the sorites paradox. In Lotte Sommerer and Elena Smirnova eds. *Nodes and Networks in Diachronic Construction Grammar*. Amsterdam: John Benjamins, 45–68.
- Givón, Talmy. 1984. *Syntax: A Functional-Typological Introduction* (Vol. 1). Amsterdam: John Benjamins.
- Goldberg, Adele E. 1991. It can't go down the chimney up: Paths and the English resultative. *Berkeley Linguistic Society* 17: 368–378.
- Goldberg, Adele E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: The University of Chicago Press.
- Goldberg, Adele E. and Ray Jackendoff. 2004. The English resultative as a family of constructions. *Language* 80: 532–568.
- Hale, Ken and Samuel J. Keyser. 1986. Some transitivity alternations in English. *Lexicon Project Working Papers* 7: 605–638.
- Hale, Ken and Samuel J. Keyser. 2002. *Prolegomenon to a Theory of Argument Structure*. Cambridge, Mass.: The MIT Press.
- Han, Ligang. 2019. A review of the major varieties of English language. *International Education Studies* 12/2: 93–99.
- Haspelmath, Martin. 1993. More on the typology of inchoative/causative verb alternations. In Bernard Comrie and Maria Polinsky eds. *Causatives and Transitivity*. Amsterdam: John Benjamins, 87–120.
- Hickey, Raymond. 2012. Standard English and standards of English. In Raymond Hickey ed. *Standards of English: Codified Varieties around the World*. Cambridge: Cambridge University Press, 1–33.
- Hilpert, Martin. 2014. *Construction Grammar and its Application to English*. Edinburgh: Edinburgh University Press.
- Höche, Silke. 2009. *Cognate Object Constructions in English: A Cognitive Linguistic Account*. Tübingen: Gunter Narr Verlag.
- Hopper, Paul J. and Sandra A. Thompson. 1980. Transitivity in grammar and discourse. *Language* 56/2: 251–299.
- Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

- Israel, Michael. 1996. The way constructions grow. In Adele E. Goldberg ed. *Conceptual Structure, Discourse and Language*. Stanford: CSLI Publications, 217–230.
- Jackendoff, Ray. 1990. *Semantic Structures*. Cambridge, Mass.: The MIT Press.
- Jackendoff, Ray. 1997. Twistin' the night away. *Language* 73/3: 534–559.
- Kim, Mija. 2010. On the time away construction: A corpus-based approach. *Linguistic Research* 27/1: 121–136.
- Kim, Jong-Bok and Jooyoung Lim. 2012. English cognate object constructions: A usage-based construction grammar approach. *English Language and Linguistics* 18: 31–55.
- Kijparnich, Nabhidh. 2011. The unergative-unaccusative split: A study of the verb *die*. *Manutsayasat Wichakan* 18/2: 107–126.
- Kudrnáčová, Naděžda. 2005. Oscillatory corporeal verbs from a semantico-syntactic perspective. *Brno Studies in English* 31/1: 35–48.
- Kuno, Susumu and Ken-ichi Takami. 2004. *Functional Constraints in Grammar. On the Unaccusative-Unergative Distinction*. Amsterdam: John Benjamins.
- La Polla, Randy J., František Kratochvíl and Alexander R. Coupe. 2011. On transitivity. *Studies in Language* 35/3: 469–491.
- Lakoff, Robin. 1973. The logic of politeness: or, minding your P's and Q's. In Claudia W. Corum, Thomas Cedric Smith-Stark and Ann Weiser eds. *Proceedings from the 9th Regional Meeting of the Chicago Linguistic Society*. Chicago: Linguistic Society, 292–305.
- Lakoff, George and Mark Johnson. 1980. *Metaphors We Live By*. Chicago: The University of Chicago Press.
- Leech, Geoffrey, Paul Rayson and Andrew Wilson. 2001. *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. London: Routledge.
- Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: The University of Chicago Press.
- Levin, Beth and Malka Rappaport Hovav. 1995. *Unaccusativity: At the Syntax-Lexical Semantics Interface*. Cambridge, Mass.: The MIT Press.
- Levinson, Stephen C. 1983. *Pragmatics*. Cambridge: Cambridge University Press.
- Macfarland, Talke. 1995. *Cognate Objects and the Argument/Adjunct Distinction in English*. Evanston, Illinois: The Northwestern University dissertation.
- Marantz, Alec. 1992. The way-construction and the semantics of direct arguments in English: A reply to Jackendoff. *Syntax and Semantics* 26: 179–188.
- Martínez Vázquez, Montserrat. 1998. *Diátesis: Alternancias oracionales en la lengua inglesa*. Huelva: Servicio de Publicaciones de la Universidad de Huelva.
- Martínez Vázquez, Montserrat. 2014. Expressive object constructions in English: A corpus-based analysis. *Revista Canaria de Estudios Ingleses* 69: 175–190.
- Massam, Diane. 1990. Cognate objects as thematic objects. *The Canadian Journal of Linguistics* 35: 161–90.
- McClure, William. 1990. A lexical semantic explanation for unaccusative mismatches. In Katarzyna Dziwirek, Patrick Farrell and Errapel Meijas-Bikandi eds. *Grammatical Relations: A Cross-Theoretical Perspective*. Stanford: CSLI Publications, 305–318.
- McColm, Dan. 2019. *A Cross-Linguistic Investigation of the Way-Construction in English, Dutch, and German*. Edinburgh: The University of Edinburgh dissertation.
- McMillan, Allan. 2006. *Labile Verbs in English: Their Meaning, Behavior and Structure*. Stockholm: The Stockholm University dissertation.
- Mittwoch, Anita. 1998. Cognate objects as reflections of Davidsonian arguments. In Susan Rothstein ed. *Events and Grammar*. Dordrecht: Kluwer, 309–332.

- Næss, Åshild. 2007. *Prototypical Transitivity*. Amsterdam: John Benjamins.
- Nakajima, Heizo. 2006. Adverbial cognate objects. *Linguistic Inquiry* 37: 647–684.
- Ogata, Takafumi. 2011. Cognate objects as categorical expressions. *Journal of Chikushi Jogakuen University and Junior College* 3: 1–14.
- Payne, Thomas. 1997. *Describing Morphosyntax*. Cambridge: Cambridge University Press.
- Peña Cervel, Sandra M. 2009. Constraints on subsumption in the caused-motion construction. *Language Sciences* 31: 740–765.
- Perek, Florent. 2018. Recent change in the productivity and schematicity of the way-construction: A distributional semantic analysis. *Corpus Linguistics and Linguistic Theory* 14: 65–97.
- Perlmutter, David. 1978. Impersonal passives and the Unaccusative Hypothesis. In Jeri J. Jaeger, Anthony C. Woodbury, Farrell Ackerman, Christine Chiarello, Orin D. Gensler, John Kingston, Eve E. Sweetser, Henry Thompson and Kenneth W. Whistler eds. *Proceedings of the 4th Annual Meeting of the Berkeley Linguistics Society*. Berkeley: Berkeley Linguistics Society, 157–190.
- Radford, Andrew. 1988. *Transformational Grammar: A First Course*. Cambridge: Cambridge University Press.
- Real Puigdollers, Cristina. 2008. The nature of cognate objects: A syntactic approach. In Sylvia Blaho, Camelia Constantinescu and Bert Le Bruyn eds. *Proceedings of ConSOLE XVI*. Leiden: Leiden University, 157–178.
- Riaubienė, Benita. 2015. *Resultative Secondary Predicates in European Languages*. Lithuania: Vilna University.
- Rivas, Elena. 1996. Construcciones de objeto interno en castellano medieval. Intento de caracterización. *Revista de Filología Románica* 13: 39–60.
- Roberge, Yves. 2002. Transitivity requirement effects and the EPP. *Paper presented at WECOL 2002*. Vancouver: Universidad de British Columbia.
- Rodríguez Adrados, Francisco. 1992. *Nueva Sintaxis del Griego Antiguo*. Madrid: Gredos.
- Sailer, Manfred. 2010. The family of English cognate object constructions. In France Stefan Müller ed. *Proceedings of the 17th International Conference on Head-Driven Phrase Structure Grammar*. Stanford: CSLI Publications, 191–211.
- Smith, Carlota S. 1970. Jespersen's 'Move and Change' class and causative verbs in English. In Mohammad Ali Jazayery, Edgar C. Polomé and Werner Winter eds. *Linguistics and Literary Studies in Honor of Archibald A. Hill* (Vol. 2). Mouton: The Hague, 101–109.
- Snell-Hornby, Mary. 1983. *Verb Descriptivity in German and English: A Contrastive Study in Semantic Fields*. Heidelberg: Carl Winter.
- Song, Jae Jung. 1996. *Causatives and Causation: A Universal-Typological Perspective*. London: Longman.
- Stockwell, Robert, Paul Schachter and Barbara Partee. 1973. *The Major Syntactic Structures of English*. New York: Holt, Rinehart and Winston.
- Sweet, Henry. 1891. *A New English Grammar. Part I: Introduction, Phonology, and Accidence*. Oxford: Clarendon.
- Talmy, Leonard. 1985. Lexicalization patterns: Semantic structure in lexical forms. In Tim Shopen ed. *Language Typology and Lexical Descriptions. Grammatical Categories and the Lexicon*, 3. Cambridge: Cambridge University Press, 57–149.
- Talmy, Leonard. 2000. *Toward a Cognitive Semantics II: Typology and Process in Concept Structuring*. Cambridge, Mass.: The MIT Press.

- Taylor, John R. 1995. *Linguistic Categorization: Prototypes in Linguistic Theory*. Oxford: Oxford University Press.
- Thalberg, Irving. 1972. *Enigmas of Agency*. London: George Allen and Unwin.
- Tsunoda, Tasaku. 1985. Remarks on transitivity. *Journal of Linguistics* 21/2: 385–396.
- van Gelderen, Elly. 2018. *The Diachrony of Verb Meaning: Aspect and Argument Structure*. London: Routledge.
- Visser, Fredericus Theodorus. 1963–1973. *A Historical Syntax of the English Language*. Leiden: J. Brill.
- Wierzbicka, Anna. 1997. *Understanding Cultures Through Their Key Words: English, Russian, Polish, German, and Japanese*. Oxford: Oxford University Press.
- Wilson, Jacob. 2019. *The Syntax and Lexical Semantics of Cognate Object Constructions*. Arizona: Arizona State University.
- WorldData.info. 2025. English speaking countries. <https://www.worlddata.info/languages/english.php> (22 February, 2025.)

Corresponding author

Beatriz Rodríguez Arrizabalaga
 University of Huelva
 Department of English Philology
 Avda. de las Fuerzas Armadas, s/n
 ES-21007 Huelva
 Spain
 E-mail: arrizaba@uhu.es

received: December 2024
 accepted: July 2025

Review of Martín Arista, Javier and Ana Elvira Ojanguren López. 2024. *Structuring Lexical Data and Digitising Dictionaries: Grammatical Theory, Language Processing and Databases in Historical Linguistics*. Leiden: Brill. 412 pp. ISBN: 978-90-04-70266-0. <https://doi.org/10.1163/9789004702660>

Silvia Saporta Tarazona
University of La Rioja / Spain

In today's world, Artificial intelligence (AI) has revolutionised our understanding of technology, which has been primarily attributed to the emergence of Large Language Models (LLMs) and their exceptional ability to perform Natural Language Processing (NLP) tasks such as text generation, summarisation, sentiment analysis or machine translation. Considering this scenario, an effective organisation of the available linguistic data is deemed imperative, especially in the case of historical languages, as datasets and databases are extensively employed by NLP applications. To accomplish this, Martín Arista and Ojanguren López's proposal establishes the underpinnings of corpus compilation through historical lexicography and lexicology from a two-pronged approach, namely, the inclusion of research procedures arising from lexical databases and language processing, and the configuration of historical dictionaries for lexicography and corpus analysis. This book constitutes a contribution to the avenue of research in the processing of historical texts with computational and artificial intelligence approaches also pursued in recent publications such as Villa and Giarda (2023), Martín Arista (2024) and Martín Arista *et al.* (2025).

Whilst Chapter 1 discusses the current state of research and underlines the major contents of the book, the ensuing chapters, which are structured in two parts titled "Lexical databases and language processing in digital historical lexicography" and "Structuring historical lexicons for lexicography and corpus analysis," delve into languages such as Old English (OE), Old Church Slavonic (OCS), Sanskrit or Greek and explore lexical processes including lemmatisation, encoding, semantic structure, lexical representation, dialectal lexicography or digitisation, among others. Thus, Chapter 2, "String similarity



measures for evaluating the lemmatisation in Old Church Slavonic,” authored by Ilia Afanasev and Olga Lyashevskaya, seeks to evaluate the most significant metrics for the lemmatisation of an OCS corpus-based dictionary, as this language has not been efficiently digitised despite its limited number of sources. Having demonstrated that the implementation of accuracy score metrics on two distinct datasets is not able to reflect central erroneous patterns, special emphasis has been placed on string similarity measures, that is to say, the Levenshtein distance (Levenshtein 1966), the Damerau-Levenshtein distance (Damerau 1964) and the Jaro-Winkler distance (Jaro 1989; Winkler 1990), which aim for an efficient, stable and scalable language model tuning.

In Chapter 3, entitled “Encoding the specificities of encyclopedias,” Alice Brenon underlines the need for a shift in the current encoding of encyclopedias, for the latter have proved to be dissimilar from conventional dictionaries as “not only do entries tend to be longer [...], they often have a deeper structure.” (p. 60). The impossibility to apply the dictionary module comprised in the encoding standard XML-TEI to works such as *La Grande Encyclopédie* has led the author to develop a new encoding scheme based on graph theory, which consistently adheres to XML-TEI, so as to fully represent the complexity entailed in encyclopedic content and ensure its accessibility to the scientific community. The subsequent chapter, “Challenges in the process of retro-digitisation of Croatian grammar books before Illyrism” by Marijana Horvat, Martina Kramarić and Ana Mihaljević, examines the main obstacles encountered in the retro-digitisation of a selection of pre-standard Croatian grammar books in an effort to design a model capable of such endeavour. Among these, encoding has emerged as a primary difficulty, for it demands multilevel annotation by means of TEI tags which are currently inexistent in the case of Illyrian grammar books. In view of this, Horvat, Kramarić and Mihaljević have elaborated a TEI Header along with a terminology index which provide “essential insights into the digitisation process of all Croatian historical documents” (p. 8) and pave the way for a comprehensive description of Croatian language history, thus becoming the most substantial contribution of this project. Ellert Thor Johannsson identifies in Chapter 5, entitled “The evolution of a dictionary of Old Norse Prose (ONP): from a collection of citations to a digital resource,” the crucial aspects for the configuration of a digital dictionary that compiles the vocabulary found in Norwegian and Icelandic medieval manuscripts. To this end, the phases required for the transition from a collection of citations to an online digital resource are painstakingly noted, which not only include the

organisation and registration of data in a lexical database that serves as the fundamental element of the digital dictionary, but also electronic applications that supplement its data with the aim of enabling users to interact with these resources and enhance their understanding of Old Norse language, literature and culture. In Chapter 6, “Agile lexicography: rapid dictionary prototyping with R Shiny, with examples from projects on Sanskrit and Tibetan,” Ligeia Lugli illustrates the potential benefits identified in Shiny, an open-source framework which develops web applications by means of R programming language (Chang *et al.* 2021). Its extra functionality was exemplified in projects involving the compilation of historical dictionaries, namely, a dictionary and thesaurus of Buddhist Sanskrit and a diachronic corpus of Tibetan verb valency (Lugli 2019; Lugli *et al.* 2022; Pagel *et al.* 2021). The simplicity and flexibility of the aforementioned tool has revealed itself as particularly well-suited for data-intensive applications, and has also facilitated the creation of new functionalities, a creative design and prototyping, and a data-pipeline while adhering to time and budget constraints. Despite the fact of having minor shortcomings related to scalability, speed or latency, it is consistently regarded as a paramount framework for historical digital lexicography.

Javier Martín Arista addresses in the seventh chapter, entitled “Interface of Old English dictionaries in database format: toward a knowledge base,” the limited presence of OE in textual and lexical databases, which mainly stems from the lack of annotation and the incompatibility of its lexical resources. With this state of affairs, an interface able to merge information from different sources is introduced, in pursuit of providing full availability and access to its linguistic data and enhancing the compilation of Old English data. This undertaking entails the digitisation of lexicographical sources and the population of the relational database, as well as the development of knowledge graphs, giving rise to a knowledge base which allows multivariable queries and ensures compatibility with NLP and AI resources. The following contribution, “Bosworth-Toller’s Anglo-Saxon Dictionary online” by Ondřej Tichý and Martin Roček (Chapter 8), sheds light on the digitisation process of the *Anglo-Saxon Dictionary*, “the only fairly comprehensive dictionary of Old English available to both experts and the public.” (p. 184). Furthermore, its most significant obstacles are also underlined, devoting special attention to the disambiguation of references, the quality of its database and the accuracy of its digital representation, as well as future updates. Notwithstanding these remaining challenges, the digital Dictionary has effectively surpassed its printed version and

constitutes a formidable lexicographical resource in the field, which will eventually develop into a high-quality dataset employed by a wide range of users.

Chapter 9, entitled “The adjective *gesælig* in Old English prose: towards the characterization of the lexical field of holiness in Old English” by Ondřej Fúsik and Alena Novotná, aims to provide a detailed portrayal of the OE lexical field of being holy, with a particular focus on the adjective (*ge*)*sælig*, so as to clarify its meaning and its existing relation with similar adjectives. To address said purpose, the lexicographic profile of this adjective has been examined by means of the *Thesaurus of Old English*, the *Bosworth-Toller Dictionary* and the *Dictionary of Old English*, although the *York-Toronto-Helsinki Parsed Corpus of Old Prose* has demonstrated to be key for the compilation of relevant data. Fúsik’s (2018) prior contributions have also been deemed indispensable to enable meaningful comparisons with the adjective *halig*, which have evidenced that (*ge*)*sælig* might be more accurately translated as ‘blessed’, and that it is primarily used in the predicative function as opposed to the former, which is eminently attributive. Despite the fact that both adjectives are found in similar genres, the authors conclude that the holiness of (*ge*)*sælig* derives from good behaviour, whereas the blessedness of *halig* originates directly from God. In Chapter 10, entitled “Cultural labels as a means of organizing semantic structure of lexemes in an explanatory synchronic historical dictionary,” Alenka Jelovšek thoroughly analyses the existing label classification for historical and specialised dictionaries (Hausmann 1989; Atkins and Rundell 2008) to present a universal typology based on their function. According to the author, encyclopedic, linguistic and register labels exhibit greater efficacy compared to encyclopedic notes, for it has been noted in the labelling of the Dictionary of the sixteenth-century Slovenian literary language, a synchronic historical dictionary detailing the period of the Slovenian Reformation. In light of this, it could be claimed that systematised labels constitute a plausible method which simplify the use of dictionaries, promote the standardisation of historical resources and provide specific meanings considering historical, ideological and textual dimensions.

Chapter 11, “Organising the lexicon by means of grammatical behaviour: the verbal class of Deprive in Old English” by Miguel Lacalle Palacios, seeks to evaluate Old English verbal lexicon, in particular, those verbs denoting the meaning of depriving, along with the constructions and alternations typically associated with them. Thus, the Role and Reference Grammar (RRG) (Foley and Van Valin 1984; Van Valin and LaPolla

1997) and the framework of verbal classes and alternations (Levin 1993) have served as the foundation of the analysis. Equally significant for the extraction of data have been the diverse textual and lexicographical sources of the study, which comprise the *Dictionary of Old English Corpus*, the *York-Toronto-Helsinki Parsed Corpus of Old English Prose* or the *Thesaurus of Old English*. The findings of the undertaking have been able to illustrate the relationship between semantics and morpho-syntactic alternations, as well as proposing four alternations and two constructions that have enabled the evaluation of the aforementioned category of OE verbs. In Chapter 12, entitled “On lemmas and dilemmas again: problems in historical dialectal lexicography,” Io Manolessou and Georgia Katsouda discuss the lemmatisation obstacles encountered in historical and dialectal lexicography, since headword selections may constitute a considerable challenge owing to the divergence of forms included in the same heading. The study has been conducted through specific instances observed in two main Greek lexicographic projects: the *Historical Dictionary of Modern Greek (ilne)* and the *Historical Dictionary of the Cappadocian Dialects*, which have provided the means for the compilation of specifically complex cases and their possible solutions. Furthermore, the criteria for headword selection has been established, although these are subjected to variation depending on the purpose of the dictionary, source availability or intended users.

Conversely, in Chapter 13, “Structuring the lexicon of Old English with syntactic principles: the role of deverbal nominalisations with aspectual and control verbs,” Ana Elvira Ojanguren López explores the association between Old English semantic and syntax through the functions of deverbal nominalisations in aspectual and control verbs, which constitutes a substantial contribution, considering that the historical evolution of the English gerund relies on the acquisition of verbal properties by this type of nominalisations (Fischer 1992). To conduct such an endeavour, the theory of RRG (Van Valin and LaPolla 1997) has been employed as the theoretical framework of the study, whereas resources such as the *York-Toronto-Helsinki Parsed Corpus of Old English Prose* or the *Dictionary of Old English* have become pivotal for data retrieval. The results lead to claim that OE syntactic configurations, including derived constructions with deverbal nominalisations, are deemed a principle capable of structuring verbal lexicon. The closing chapter (Chapter 14), entitled “Assessing lexicographic obsolescence and historical frequency indicators in word entries in the OED: a corpus study of historical *-some* adjectival derivatives” by Chris Smith,

implements a diachronic perspective to examine the field of obsolescence and low-frequency words in lexicographic databases such as the *Oxford English Dictionary* (OED). To this end, a systematic methodology is proposed to assess frequency labelling in a specific set of -some adjectival derivatives, which reveals the wide variability of obsolescence and relative frequency, and contributes to the study of lexicographic labelling and diachronic lexical competition.

Individual chapters have both strengths and limitations that deserve some comment. While the methodology of Chapter 2 is thorough, the study could benefit from more contextual examples illustrating lemmatisation problems. Moreover, some interpretation of the relationship of the metrics to general linguistic phenomena would strengthen the conclusions. The graph theory has revealed itself as an innovative approach, yet real-world examples of encoding problems and decisions would increase readability. Plus, the distinction between dictionaries and encyclopedias could be further clarified through structural diagrams, instead of relying on verbal descriptions only. Chapter 4 provides a strong contextual background, although the technical implementation of TEI headers draws more attention than the evaluation of the impact of the digitisation process. In spite of this, the grammar book comparison table is deemed excellent, although employing case studies to examine particular problematic elements could enhance comprehension among general audiences. Conversely, the author in Chapter 5 painstakingly illustrates database structure along with the evolution from analog to digital formats, but more in-depth critical examination of how lexicographical decisions affected the digitisation process would be highly beneficial. Additionally, more analysis of user experience with the final digital product could help to assess the impact of digitisation.

Chapter 6 possesses a strong practical focus and provides convincing examples, although more emphasis should be made on the limitations of Shiny for production environments. The comparison between different development approaches is particularly valuable, yet more quantitative metrics on performance would strengthen the argument. Chapter 7 constitutes a significant contribution to Old English lexicography through the development of the *Interface of Old English Dictionaries* (IOED), a component of the *Knowledge Base of Old English* (KBOE). Moreover, the methodology of converting relational databases to knowledge graphs represents an innovative approach to historical lexicography, whereas the combination of type analysis (lemmas) with token analysis (inflectional forms) provides a comprehensive resource for both historical corpus

linguists and lexicographers. The chapter excels in demonstrating how inconsistencies across dictionaries (regarding headword spelling, vowel quantity, etc.) can be resolved through standardisation and normalisation in a database format. In addition to this, the graphs generated from edge lists demonstrate the potential for more efficient complex queries. Chapter 8 effectively documents the transformation process from print to digital dictionary by incorporating an excellent historical context. Nonetheless, alternatives to the chosen encoding approaches might have been discussed, and the influence of user needs on technical decisions could have received more focus.

Chapter 9 presents a thorough corpus-based investigation of the Old English adjective (*ge*)*sælig* and its position within the lexical field of holiness. While the study acknowledges limitations due to the restricted corpus (mainly religious texts), it makes a convincing case that translating (*ge*)*sælig* as ‘blessed’ rather than ‘happy’ may be more appropriate in many contexts. Chapter 10 proposes a three-class labeling system (encyclopedic, linguistic, register) and therefore contributes to bridging the gaps in previous classifications. However, the chapter could benefit from more concrete examples demonstrating how the proposed structure improves digital interoperability beyond theoretical frameworks. Chapter 11 applies the framework of verb classes and alternations with Role and Reference Grammar to analyse Old English verbs of depriving and convincingly demonstrates that grammatical behaviour constitutes a sounder basis for lexical organisation than meaning definitions alone. Plus, the chapter succeeds in presenting a rigorous approach that could be applied to other verbal classes in historical English. As it can be noted, Chapter 12 presents a strong analysis of lemmatisation issues in Greek lexicography with excellent discussion of methodology. Whilst the examples are detailed and well-chosen, a visual depiction of the decision tree for selecting headwords would enhance readability.

Chapter 13 analyses deverbal nominalisations with aspectual and control verbs in Old English. The methodology combining lexical database analysis with corpus linguistics offers insights into both synchronic and diachronic aspects of the language and demonstrates that nominal linked predications display semantic and syntactic configurations parallel to verbal linked predications. Furthermore, the analysis of the macrorole transitivity of nominal linked predications represents a particularly original contribution to the field. The author convincingly argues that the syntactic configurations of Old English can serve as a principled basis for structuring the verbal lexicon. Chapter

14 constitutes an impressive quantitative analysis of *-some* adjectives, although the methodology section could have connected the corpus testing techniques with the theoretical framework more clearly. In spite of the visual presentation of data in figures supporting the arguments about frequency patterns, more discussion of the implications for ongoing OED revisions would help general audiences to understand the connection between lexicographical theory and practice.

Overall, *Structuring Lexical Data and Digitising Dictionaries* represents a significant advance in historical lexicography and computational linguistics, particularly for diachronic English studies. The interdisciplinary approach of the volume, which combines traditional philological methods with cutting-edge computational techniques, offers innovative solutions to long-standing problems met when accessing structuring, and analysing historical lexical data. Across thirteen chapters, the contributors demonstrate how relational databases, knowledge graphs, corpus-based semantic analysis, and syntactic frameworks can organise lexical information from various sources. The book particularly excels in addressing practical problems of standardisation across dictionaries and corpora, given that it proposes methodologies for semantic field analysis, and establishes principled approaches to verb classification that underline the relationship between semantics and syntax. While some theoretical frameworks might be applied too rigidly to historical data, and certain computational methodologies would benefit from more rigorous evaluation, the book as a whole significantly improves our understanding of how to transform traditional lexicographical resources into structured digital datasets. This volume will undoubtedly become a reference work for future projects in digital historical lexicography.

REFERENCES

- Atkins, B. T. Sue and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Chang, Winston, Joe Cheng, J. J. Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert and Barbara Borges. 2021. *Shiny*: <https://CRAN.R-project.org/package=shiny>
- Damerau, Frederick J. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM* 7/3: 171–176.
- Fischer, Olga. 1992. Syntax. In Norman Blake ed. *The Cambridge History of the English Language II. 1066- 1476*. Cambridge: Cambridge University Press, 207–407.
- Foley, William and Robert Van Valin. 1984. *Functional Syntax and Universal Grammar*. Cambridge: Cambridge University Press.
- Fúsik, Ondřej. 2018. *Old English Prose Adjectives Meaning ‘holy’: Towards a Characterization of a Lexical Field*. Prague: Charles University dissertation.

- Hausmann, Franz Josef. 1989. Die Markierung in einem allgemeinen einsprachigen Wörterbuch: eine Übersicht. In Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand and Ladislav Zgusta, eds. *Wörterbücher: Ein Internationales Handbuch zur Lexikographie*. Berlin: Walter de Gruyter, 649–657.
- Jaro, Matthew A. 1989. Advances in record linkage methodology as applied to the 1985 Census of Tampa Florida. *Journal of the American Statistical Association* 84: 414–420.
- Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10/8: 707–710.
- Levin, Beth. 1993. *English Verb Classes and Alternations*. Chicago: University of Chicago Press.
- Lugli, Ligeia. 2019. Smart lexicography for low-resource languages: Lessons learned from Sanskrit and Tibetan. In Iztok Kosem, Tanara Zingano Kuhn, Margarita Correia, José Pedro Ferreira, Maarten Jansen, Isabel Pereira, Jelena Kallas, Miloš Jakubiček, Simon Krek and Carole Tiberius eds. *Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the eLex 2019 Conference*. Sintra, 198–212.
- Lugli, Ligeia, Matej Martinc, Andraž Pelicon and Senja Pollak. 2022. Embeddings models for Buddhist Sanskrit. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk and Stelios Piperidis eds. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, 3861–3871.
- Martín Arista, Javier. 2024. Toward a Universal Dependencies Treebank of Old English: Representing the morphological relatedness of un-derivatives. *Languages* 9/3: 76.
- Martín Arista, Javier, Ana Elvira Ojanguren López and Sara Domínguez Barragán. 2025. Universal Dependencies annotation of Old English with spaCy and MobileBERT: Evaluation and perspectives. *Procesamiento del Lenguaje Natural* 75: 253–262.
- Pagel, Ulrich, Edward Garrett, Ligeia Lugli and Christian Faggionato. 2021. *A Visual Dictionary of Tibetan Verb Valency*. Mangalam Research Institute. <https://doi.org/10.5281/zenodo.5596064>
- Van Valin, Robert and Randy J. LaPolla. 1997. *Syntax: Structure, Meaning and Function*. Cambridge: Cambridge University Press.
- Villa, Luca Brigada and Martina Giarda. 2023. Using modern languages to parse ancient ones: A test on Old English. *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2023)*. Dubrovnik: Association for Computational Linguistics, 30–41.
- Winkler, William Erwin. 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods*. American Statistical Association: 354–359.

Reviewed by

Silvia Saporta Tarazona
University of La Rioja
Department of Modern Philologies
C/ San José de Calasanz, 33
E-26004, Logroño, La Rioja
Spain
e-mail: silvia.saporta@unirioja.es

Review of Ljubica Leone. *Composite Predicates in Late Modern English* (Routledge Focus on Linguistics). 2024. London: Routledge. 92 pp. ISBN 9781003410256.
<https://doi.org/10.4324/9781003410256>

Harumi Tanabe
Seikei University (retired) / Japan

Since the publication of Brinton and Akimoto's pioneering work, *Collocational and Idiomatic Aspects of Composite Predicates in the History of English* in 1999, increasing interest in composite predicates (CPs),¹ verbal structures consisting of a verb + an indefinite article + a deverbal noun or eventive object (Quirk *et al.* 1985: 750–752), has been consistently shown in various studies: Claridge (2000) explores the *Lampeter Corpus of Early Modern English Tracts* in 1640–1740, Moralejo Gárate (2003) examines the Middle English period of the *Helsinki Corpus*, supplementing Kytö (1999), Bergs (2005) includes a chapter concerning the light verb constructions in the Paston Letters, Wang (2019) analyses the speech-focused text in *A Corpus of English Dialogues in 1560–1760* and in even more recent work Berlage (2025) investigates present-day English CPs. In line with this trend, Leone's monograph *Composite Predicates in Late Modern English* (2024) is highly welcomed, as it studies CPs in Late Modern English (LModE) with a view to portraying ongoing linguistic change through a historical survey of the CPs in this period. Leone intends to fill the gap in research on CPs in the LModE period on which scholarly work in CPs has been so far sparse, except for Akimoto (1999). Besides, she aims “to test whether and the extent to which these verbs were affected by similar paths of development that characterized other multi-word verbs of the LModE time as examined in other works” (Leone 2024: 12). Having already published another volume, *Multi-word Verbs in the Late Modern English Period (1750–1850)*, which examines phrasal verbs, prepositional verbs and phrasal-prepositional verbs in the LModE part of

¹ This verbal construction has been variously referred to since Poutsma regarded it as one type of group-verb as early as in 1926. The term composite predicate was first used by Cattell (1984). Other terms are complex verb (Nickel 1968), expanded predicates (Algeo 1995), verbo-nominal structure (Akimoto 1989, 1999), verbo-nominal combination (Claridge 2000), just to name a few.



the *Old Bailey Corpus (LModE-OBC)*, she has chosen the same corpus for her study of CPs to facilitate comparison of the results between CPs and other multi-word verbs.

Verbs used in CPs such as *do*, *give*, *have*, or *make* are known as light verbs as they have undergone semantic bleaching with more emphasis on the nouns to express the meaning of the whole unit. An indefinite article preceding a noun is optional or sometimes substituted with other determiners such as *the*, *this*, *that*, *some*, *any*, *no*, or personal pronouns. Leone's corpus-based study investigates CPs appearing in her self-edited *LModE-OBC*, an extract from Late Modern English texts (1750-1850) drawn from the *Old Bailey Corpus*, which is a collection of trial depositions and dialogues recorded in London's Central Criminal Court from 1674 to 1913.

Chapter 1 "Composite Predicates in 1750-1850" introduces the structure, linguistic features, a brief historical outline of CPs in general, and explains the compilation process and the details of the corpus used in the study. In Section 1.5 "Method", selection criteria for light verbs and deverbal nouns or eventive objects are briefly explained. Following Brinton and Akimoto (1999), the light verbs *do*, *give*, *have*, *make*, and *take* are chosen and as the nominal element i) a deverbal noun of the same form as a simple verb, ii) a deverbal noun derived by suffixation and iii) a noun etymologically related to the simple verb (p. 10). To ensure the etymological relationship between nouns and verbs, all the instances of nouns are checked against the OED to see if they have the relevant verbal form. If they do, regardless of their historical status or semantic correspondences, they are included as an element of CP (pp. 10–11). As for the statistical methods, the author calculates the raw and normalized frequencies (NFs), percentages, type/token ratio, log-likelihood score and logistic regression to assess the degree of productivity and the statistical significance (pp. 11–12).

Chapter 2 "History" outlines a more detailed history of CPs by introducing the previous studies, focusing on the diachronic development of idiomaticity (pp. 14–25). In Old English, there are some combinations regarded as CPs but they are not fully established as a unit equivalent to what we call CPs now. In Middle English, although the frequencies are still not so high, CPs are certainly emerging. In Early Modern English, Hiltunen (1999) and Kytö (1999) conducted thorough surveys in the *Helsinki Corpus* and found *have* and *make* to be most frequent and productive base verbs. In *A Corpus of English Dialogues*, the corpus of drama and court trials from 1640-1740, Wang (2019)

examined the frequency, lexical productivity and syntactic features of the light verbs in drama and trial texts 1640-1740.

Leone regards the LModE period as a transition from a synthetic to a more analytic language, and quoting Akimoto (1999), who examined CPs in the eighteenth and nineteenth century texts and the OED citations, she highlights, as Akimoto (1999: 225) finds, a remarkable tendency towards idiomatization (p. 22).

Chapter 3 “Linguistic features” displays the results of the survey of CPs. The overall NF of CPs in this period, 3.01 NF per 1000 words, is revealed to be increasing compared with the rate in former periods (p. 25), but within the period of 1750-1850 the rate stays almost stable with a decrease in the 1810s and a quick recovery in the 1830s (p. 27, Figure 3.1). The log-likelihood score of the increase from 3.21 NF in the 1750s to 3.45 NF in the 1830s shows this is not statistically significant. Furthermore, to check the relationship between time and frequency, the author, saying she will use logistic regression, in fact applies linear regression (p. 28). As a result, the line in Figure 3.2 (p. 28) indicates a gradual decrease in frequency, but it is also found to be statistically insignificant. Leone’s quantitative analysis lacks sufficient discussion, but she seems to imply these results can be evidence of the stability of CPs in this period. To prove the data’s stability, however, she needs appropriate statistical procedures.

The productivity of CPs is evaluated, based on the variation in light verbs and nouns. It is found that the higher frequency verbs, *have*, *give*, and *make* tend to occur with a wider range of nouns, while the lower frequency verbs, *do* and *take*, combine with a rather limited number of nouns (p. 34).

Chapter 4 “Composite predicates between stability and change” discusses mainly morphosyntactic features of CPs, providing a general portrait of syntactic forms showing stability in LModE in comparison with those of the EModE and PDE periods, but some innovative aspects are observed at the same time (pp. 41–42). By analyzing the determiners and internal modification patterns of CPs, Leone reveals that singular nouns, especially with zero article, overwhelmingly outnumber the plural counterparts in her *LModE-OBC*, showing a substantial increase compared with the data in the EModE period obtained from Kytö (1999). This may support the trend toward lexicalization and grammaticalization together with appearance of high-frequency fixed expressions such as *make a great noise* (pp. 49–50). This chapter concludes by proposing a cline of

semantic compositionality in CPs, ranging from literal (e.g. *have account*) to idiomatic (e.g. *make haste*) (pp. 64–66).

Chapter 5 “Process of change” further discusses how change took place in terms of grammaticalization, lexicalization and idiomatization. Drawing on Akimoto’s (1989, 1999) four-stage theory of idiomatization² which evolves from free phrasal combinations of components to lexicalization, to syntactic and semantic reanalysis and subsequently to idiomatic combinations with syntactic fixity (p. 55), Leone places CPs in the middle of the lexicalization cline, with reference to the criteria that can judge the degree of internal cohesion in CPs. Therefore, diachronically, as nouns in verbal units lose determiners or modification and plural forms become rarer, then the more lexicalized and idiomatized the units become as CPs (p. 56). Another criterion for lexicalization is phraseological variation and layering of prepositions. Occasionally CPs appear in different patterns such as V+(Det)+N (e.g. *have any reason to do*) or V+(Det)+N+Prep (e.g. *have any reason for doing*), while CPs can take various prepositions in the same meaning (e.g. *make observation on/upon/of/about*). CPs with these variations and layering signal incomplete lexicalization (p. 60). This leads the author to view LModE as a period closer to earlier periods. At the same time, innovative aspects, as the author considers, are seen in introduction of newly coined CPs such as *take warning*, *take opinion*, which are not recorded in Kytö (1999) and ARCHER before 1750 (p. 62),³ and also in semantic changes that some CPs have experienced due to the loss of the definite article or other determiners (p. 64).

The final chapter briefly summarizes the findings, followed by Appendix, References and Index. Regrettably, Appendix, the three-column list of CPs appearing in the *LModE-OBC*, is typographically flawed, with the final four lines in each column, probably intended to be on one page, printed on the next. As a result, the list is hard to decipher.

Throughout the discussion in the volume, the author tries to envisage a broader picture of CP change, situating it on a scale from synthesis to analyticity and applying the frameworks of lexicalization, grammaticalization and idiomatization. Her wide

² Traugott (1999) proposes three stages rather than four in a modified version of this theory.

³ Rodríguez-Puente (2025: 6) points out that earlier occurrences of the new coinages are found in OED (e.g. *make discussion*, *take warning*, *make effort*).

perspective is commendable, though a higher degree of precision would strengthen certain aspects of the analysis.

One of the points where more accuracy is expected, as Rodríguez-Puente (2025: 2–3) has pointed out, concerns the definition of CPs. The task of defining a CP may seem straightforward if CPs are confined to a verbal sequence of a verb plus a deverbal noun either with zero derivation (e.g. *take a walk*), or suffixation (e.g. *make comparison*), following simply formal criteria. However, as Leone, like Quirk *et al.* (1985: 750–752), Kytö (1999) and others, includes eventive objects as nominal elements,⁴ and applies semantic criteria, some ambiguity arises. The problem of differentiation between CPs and free combinations naturally emerges, too. Leone appears to pay limited attention to this issue, as several examples from the *LModE-OBC* quoted in the volume are unrelated to CPs ((32) *I have made all my friends my foes* (51) is in a different construction from a CP and (8) *I ... took the place at the public-house* (57) is a literal free combination). A purely formal approach, as taken by Algeo (1995), Claridge (2000) and others, might have yielded clearer results. However, if CPs are defined as constructions semantically equivalent to a simple verb with the same or similar meaning, more verbal units can be taken into consideration. This is a merit in dealing with earlier texts where the frequency of CPs is low. Leone's concise discussion of noun selection criteria leaves room for misinterpretation: the readers may be perplexed to see *dinner*, *air*, *garden*, or *time*⁵ treated as CP nouns. A more detailed discussion would have mitigated such confusion.

Another point to note is in cross-corpus comparison. The period she focuses on is only one hundred years, limited to 1750–1850. Therefore, to compare the frequencies from her period and others, she draws on the results from previous studies. But this may lead to inaccuracy. For instance, Leone compares the NF of the CPs in the *Lampeter Corpus* 1640–1740, 1.7 NF (Claridge 2000: 108, 178–179), with her results 3.01 NF in 1750–1850, claiming that “the LMod period is characterized by increasing use” (p. 25). Claridge takes up only deverbal and suffixed nouns while Leone includes eventive objects. Conversely, Claridge (2000: 40, 76–77) includes verb+prepositional phrase units

⁴ Leone follows Kytö (1999: 169) in selecting nominal elements in CPs and checking the etymological information in the OED, but in checking all the nouns she is more rigorous than Kytö, who checked the OED only in cases “intuition and the information in ordinary dictionaries are insufficient” (Kytö 1999: 206).

⁵ Rodríguez-Puente (2025: 3) considers these nouns are not suitable as an element of a CP as she does not admit nouns merely etymologically or semantically related to the verb. Opinion may be divided on this point, but as she claims, it would be better to limit the nouns to those having abstract meaning, excluding *memorandum* etc.

(Group III)) and 14 more light verbs such as *come, lay, set, put* (Claridge 2000: 120–122, Table 6.5), not present in Leone’s selection;⁶ and the *Lampeter Corpus* examined by Claridge, as Leone is well aware (p. 26), consists of religious, political, economic and scientific tracts, more formal in style than the spoken court depositions of *OBC* Leone studies. Wang’s (2019: 46) findings indicate that the drama section has higher frequencies and more variety of units than the trial section, probably because of more limited contexts in trials, clearly indicating there is a difference in the use of CPs depending on text type. All in all, comparing the NFs might end up being an imprecise enterprise.

Furthermore, the connection between the LModE period and the next period needs more exploration. The author frequently refers to previous studies in earlier times to discuss the stability of CPs in the LModE period, but the situation of CPs after 1850 is not clear enough as there is not much reference to the previous works. In fact, to demonstrate that the CPs found in the *LModE-OBC* exhibit stable features, Leone queries ARCHER and quotes similar examples in late twentieth century from it (p. 41). As ARCHER is a multi-genred corpus, an examination of the 1850-1913 part of *OBC* applying consistent criteria for text type and the definition of CPs would provide a more coherent continuation.

The reasons why the period for the survey of the *LModE-OBC* is limited to 1750-1850 are the continuity of the period from Claridge (2000) and matching the period for her study on multi-word verbs (Leone 2023). While this is understandable for comparing the results, a more detailed survey of the entire *OBC* up to early twentieth century would be more beneficial for a deeper investigation of the diachronic development. In addition, there are repeated statements that CPs in LModE have similar tendencies to those observed in multi-word verbs in Leone (2023), but without providing details of cases in multi-word verbs. Therefore, combining Leone’s two volumes on CPs and multi-word verbs into a single volume—around 160 pages—might also offer readers a more comprehensive and accessible reference.

Finally, Leone acknowledges the possibility of a declining trend in the LModE period up to 1810s (p. 67), in agreement with Akimoto (1999: 208, 215, Table 7.1), who says the CPs are more numerous in the eighteenth than in the nineteenth century, and Claridge (2000: 178), who reports a slight decline in CP frequency during 1640-1740. In

⁶ Although Claridge (2000: 122) includes Group III verb-nominal combinations, the frequency is 13 only with the base verb *take*.

addition, Berlage (2025: 189–192) finds that the specific CPs she has chosen show a decrease in frequency during the nineteenth and twentieth centuries. All these findings indicate the need for a fresh insight into the chronological change of CPs in LModE, choosing the most suitable corpus under consistent conditions and a more precise framework. To arrive at this goal, Leone’s volume certainly makes a timely and meaningful contribution to our understanding of the linguistic features, processes and mechanisms underlying the historical development of CPs, and it provides a solid foundation for future studies on the evolution of multi-word verb constructions in the history of English.

REFERENCES

- Akimoto, Minoji. 1989. *A Study of Verbo-Nominal Structures in English*. Tokyo: Shinozaki Shorin.
- Akimoto, Minoji. 1999. Collocations and idioms in Late Modern English. In Laurel J. Brinton and Minoji Akimoto eds. *Collocational and Idiomatic Aspects of Composite Predicates in the History of English*. Amsterdam: John Benjamins, 207–238.
- Algeo, John. 1995. Have a look at the expanded predicate. In Bas Aarts and Charles F. Meyer eds. *The Verb in Contemporary English: Theory and Description*. Cambridge: Cambridge University Press, 203–217.
- Bergs, Alexander. 2005. *Social Networks and Historical Sociolinguistics: Studies of Morphosyntactic Variation in the Paston Letters (1421-1503)*. Berlin: Mouton de Gruyter.
- Berlage, Eva. 2025. *Composite Predicates in English: Processes of Specialization*. Cambridge: Cambridge University Press.
- Brinton, Laurel J. and Minoji Akimoto eds. 1999. *Collocational and Idiomatic Aspects of Composite Predicates in the History of English*. Amsterdam: John Benjamins.
- Cattel, Ray. 1984. *Composite Predicates in English*. Sydney: Academic Press.
- Claridge, Claudia. 2000. *Multi-word Verbs in Early Modern English: A Corpus-based Study*. Amsterdam: Rodopi.
- Hiltunen, Risto. 1999. Verbal phrases and phrasal verbs in Early Modern English. In Laurel J. Brinton and Minoji Akimoto eds. *Collocational and Idiomatic Aspects of Composite Predicates in the History of English*. Amsterdam: John Benjamins, 133–166.
- Kytö, Merja. 1999. Collocational and idiomatic aspects of verbs in Early Modern English. In Laurel J. Brinton and Minoji Akimoto eds. *Collocational and Idiomatic Aspects of Composite Predicates in the History of English*. Amsterdam: John Benjamins, 167–206.
- Leone, Ljubica. 2023. *Multi-word Verbs in the Late Modern English Period (1750-1850): A Corpus-based Study*. Muenchen: Lincom.
- Moralejo Gárate, Teresa. 2003. *Composite Predicates in Middle English*. Muenchen: Lincom.

- Nickel, Gerhard. 1968. Complex verbal structures in English. *International Review of Applied Linguistics* 6/1: 1–21.
- Oxford English Dictionary Online*. 2025. Oxford: Oxford University Press.
<https://www.oed.com/>
- Poutsma, Hendrik. 1926. *A Grammar of Late Modern English, Part II: The Parts of Speech, Section II: The Verb and Particles*. Groningen: P. Noordhoff.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
- Rodríguez-Puente, Paula. 2025. Review of Ljubica Leone, *Composite Predicates in Late Modern English* (Routledge Focus on Linguistics). London: Routledge, 2024. Pp. vi + 84. ISBN 9781032524887. *English Language and Linguistics*: 1–7.
- Traugott, Elizabeth Closs. 1999. A historical overview of complex predicate types. In Laurel J. Brinton and Minoji Akimoto eds. *Collocational and Idiomatic Aspects of Composite Predicates in the History of English*. Amsterdam: John Benjamins, 239–260.
- Wang, Ying. 2019. A corpus-based study of composite predicates in Early Modern English dialogue. *Journal of Historical Pragmatics* 20/1: 20–30.

Reviewed by
 Harumi Tanabe
 Seikei University
 Department of English
 3-3-1 Kichijoji-Kitamachi
 180-8633, Musashino-shi, Tokyo
 e-mail: harumi.tanabe@mac.com