Ferraresi, A. and S. Bernardini (2023) 'Comparing Collocations in Translated and Learner Language: In
Search of a Method'. *International Journal of Learner Corpus Research*, 9(1): 125–153.

COMPARING COLLOCATIONS IN TRANSLATED AND LEARNER LANGUAGE: IN SEARCH OF A METHOD

Adriano Ferraresi and Silvia Bernardini

University of Bologna

This paper compares use of collocations by Italian learners writing in and translating into English, conceptualising the two tasks as different modes of constrained language production and adopting Halverson's (2017) Revised Gravitational Pull hypothesis as a theoretical model. A particular focus is placed on identifying a method for comparing datasets containing translations and essays, assembled opportunistically and varying in size and structure. The study shows that lexical association scores for dependency-defined word pairs are significantly higher in translations than essays. A qualitative analysis of a subset of collocations shared and unique to either mode shows that the former set features more collocations with direct cross-linguistic links (connectivity), and that the source/first language seems to affect both modes similarly. We tentatively conclude that second/target language salience effects are more visible in translation than second language use, while connectivity and source language salience affect both modes of bilingual processing similarly, regardless of the mediation variable.

Keywords: constrained language, Gravitational Pull Hypothesis, collocation, lexical association measures, bootstrapping

This is a post-print version of:
Ferraresi, A. and S. Bernardini (2023) 'Comparing Collocations in Translated and Learner Language: In Search of a Method'. *International Journal of Learner Corpus Research*, 9(1): 125–153.
https://doi.org/10.1075/ijlcr.22012.fer

1. Introduction

Among the objects of enquiry of corpus linguistics, collocations have constantly enjoyed a privileged status. Sinclair (1991) describes the descriptive task of linguists as "the identification of the regular and typical [word] associations", leading to the description of "integrated sense-structure complex[es]". Such complexes have been hypothesised to constitute the building blocks of our language competence because

> [a]s a word is acquired through encounters with it in speech and writing, it becomes cumulatively loaded with the contexts and co-texts in which it is encountered, and our knowledge of it includes the fact that it co-occurs with certain other words in certain kinds of context. (Hoey, 2005: 8)

This view has found support from psycholinguistics and neurolinguistics (cf. Ellis, Simpson-Vlach, and Maynard 2008; Durrant and Siyanova-Chanturia, 2015; Crossley, Salsbury and Mcnamara, 2015), further strengthening interest in the study of collocation.

Studies approaching collocation from the perspective of text-linguistic variation aim to find out how text varieties differ in their use of collocations. The learner language variety and, to a more limited extent, the translated language variety have attracted the interest of researchers who have attempted to understand if and how bilingual groups differ from first language (L1) or target language (TL) speakers in their use of collocations. More recently a unifying hypothesis has been formulated which encourages researchers to view translation and learner language as "constrained" language varieties, characterised by bilingual language activation but differing in other ways (Lanstyak and Heltái, 2012). Research in this field aims to understand which features are shared and which are unique to either variety.

Against this background, the present contribution investigates use of collocations in English translations and essays by Italian students of translation. In Section 2 we review the different research strands that this work draws from; we then introduce the datasets and the exploratory method employed (Section 3), describe our results (Section 4) and conclude by reflecting on the implications of our work for corpus-based translation studies and learner corpus research (Section 5).

2. Phraseological insights on learner language, translation and constrained language

2.1 Phraseology in learner language

Language learners' difficulties with collocations are well-known, at least since Pawley and Syder's (1983) account of native speakers' communicative competence as relying on the "puzzling" abilities of native-like selection and native-like fluency. By contrast, language learners may erroneously assume "that an element in the expression may be varied […], when in fact the variation (if any) allowed in nativelike usage is much more restricted" (ibid.: 215).

Research on learner language has investigated this hypothesis empirically, confirming that learners use fewer collocations than native speakers, both in terms of quantity and variety (see e.g., Laufer and Waldman, 2011). Extensive empirical evidence has been accumulated about factors affecting collocation use. Nesselhauf (2005) finds that German learners of English use more "congruent" collocations, i.e., collocations

which can be produced by translating word-for-word an existing German expression (ibid.: 221), than non-congruent ones, and that errors are twice as likely to affect non-congruent than congruent collocations. When appropriate non-congruent collocations are produced, these are likely to be either very frequent, such as *play a part* or *have a chat*, or semi-idiomatic, such as *fall victim* and *come into existence* (ibid.: 224-225). Thus, factors related to language contact and factors related to the nature of collocations and their constituent words both play a role in learners' phraseological choices.

The central role of contact (or transfer, or cross-linguistic influence) in learners' use of collocations has long been recognised (Kellerman, 1977). In her study of intensifiers in learner writing, Granger (1998) finds that French learners of English use fewer creative and fewer "stereotyped" collocations (very salient phrases like *acutely aware*). When they do use a stereotyped collocation, however, this has a direct translational equivalent in French. Paquot (2014) also observes contact effects in the production of lexical bundles in English by French-speaking learners. Transfer occurs at multiple levels, including collocational and colligational preferences, syntactic constructions, discourse functions and frequency of the hypothesised equivalent(s) in French. In a pseudo-longitudinal study comparing the performance of Hebrew learners of English at three proficiency levels, Laufer and Waldman (2011) detect the influence of Hebrew in about half of the erroneous collocation types produced by learners at all proficiency levels, further confirming the role of contact in terms of both size of the effect on learner performance, and development path. Contact effects have also been suggested to interact with other subject- and population-related factors, such as personal or group attitudes and teaching traditions (Wang 2016).

Alongside contact, the relationship between the frequency and the *salience* of the target expressions has been shown to affect learners' use of collocations. We use the term salience here as a short-hand for "[t]he distinctiveness, or prominence, or *salience*, or memorability, or availability of lexical items" (Kjellmer 1984: 163, our italics). Learners are found to struggle with the complex grammatical and lexical patterning of common words such as delexical verbs (Altenberg and Granger, 2001; Wang, 2016), and to produce a smaller set of less restricted, less salient collocations than native speakers (Granger, 1998; Källkvist, 1998). These collocations have memorably been referred to as learners' phrasal or phraseological "teddy bears" (Ellis 2012; Hasselgård 2019), extending Hasselgren's (1994) notion of lexical teddy bears. Using similar methods, Durrant and Schmitt (2009) and Granger and Bestgen (2014) try to disentangle the effects of frequency and salience. Native speakers and learners at a more advanced level rely more on lower-frequency, strongly associated collocations, such as those selected by the Mutual Information (MI) association measure, than learners in general (in Durrant and Schmitt's study) and intermediate learners (in Granger and Bestgen's study). Learners instead prioritise frequency over salience, as shown by greater reliance on collocations with high t-score values. These studies confirm psycholinguistic evidence highlighting the processing advantage of salience for native speakers only (Ellis et al., 2008).

One last word of caution is in order. While research has convincingly shown that second language (L2) learners use collocations differently from native speakers, one should refrain from equating distinctiveness and deficit. Bilingualism and acquisition scholars encourage us to see "the interaction between multiple languages in the speaker's mind as a natural and ongoing process", which explains "why multilinguals may perform differently from monolinguals in all of their languages, including the L1" (Jarvis and

Pavlenko, 2008: 17). An exclusive focus on native speaker performance as a yardstick against which to evaluate L2 use has indeed been criticised as "falling prey to the 'comparative fallacy' in that it fails to recognise learner language as a variety in its own right" (Callies, 2015: 49), and failing to recognise that learners have their own "wider agenda, one in which, on some occasions, a perfectly nativelike performance may be of relatively little importance" (Wray, 2002: 212).

2.2 Phraseology in translated language

Use of phraseology in translation has been approached from different angles. Using a parallel corpus of literary source texts (STs) in German and their targets (TTs) in English, together with general reference corpora of the source language (SL) and target language, Kenny (2001) investigates creativity and conventionality at the lexical and phraseological levels. She finds that creative collocations tend to be normalised through the choice of more standard options in the TL (ibid.: 208). Dayrell (2007) focuses on frequent words and their collocational patterns in a comparable corpus of narrative texts in original and translated Brazilian Portuguese. Translated texts are found "to draw heavily on a small number of collocates" (ibid.: 395), pointing, as in the case of Kenny (2001) to normalisation and conventionality. The monolingual comparable corpus approach is also used by Jantunen (2004) in his study of collocations of Finnish degree modifiers. He carries out a three-way comparison that includes non-translated Finnish, Finnish translated from English, and Finnish translated from several languages. He observes both SL independent tendencies (overuse of degree modifiers) and SL-derived lexical patterning.

Two recent studies look at collocations in translation from Chinese into English (Feng 2020) and in simultaneous interpreting from Russian into English (Dayter 2020). Feng's results confirm the preference for free combinations and high-frequency collocations, as well as the greater reliance on a smaller set of collocations in translated than non-translated texts. Dayter's less clearcut results are instead explained in the light of Shlesinger's (1989) equalizing universal, whereby interpreted speeches would converge toward an unmarked middle ground along the oral-literate continuum. While one might agree with Dayter that these findings are specific to simultaneous interpreting as a mediation modality, they also align, at a general level, with the less heightened sensitivity to register conventions observed in translation and L2 production (see Section 2.3 below).

Feng's (2020) work controversially construes professional translators working into their L2 as "essentially a special group of (advanced) L2 learners" (Feng 2020: 44), and explicitly interprets differences with respect to non-translated texts as a sign of poor performance. This prescriptive approach would raise strong objections within the translation studies community. Yet there is no doubt that the corpus-based analysis of collocations in translation has much in common with the corresponding work on learner language both in terms of methods and results. The distinctive features that set apart translated from non-translated texts in the same language and the influence of the ST/SL are obvious commonalities, while differences concern in particular the availability of a constraining ST and the fact that translators may translate either into their L1 or L2.

2.3 The constrained language hypothesis

Given the similarities between the research objects described in 2.1 and 2.2, the hypothesis has been recently made that L2 production and translation can be fruitfully construed as language activities in which several linguistic systems become simultaneously activated (Grosjean, 2013). Since L2 and translated language might diverge in similar ways from comparable native, non-translated language, it makes sense to bring the two together within the same framework in the pursuit of higher-order generalisations, or "universals of constrained communication" (Lanstyak and Heltái, 2012).

Studies of constrained communication address bi/multilingual language use and include translation as a mode of language contact. Among these, Kolehmainen, Meriläinen, and Riionheimo (2014) review evidence that interlingual reduction, operationalized as lower frequency of TL items with no obvious SL equivalents, has been observed in language contact, L2 acquisition and translation. In their study contrasting native, non-native and translated English, Rabinovich, Nisioi, Ordan, and Wintner (2016) find that non-native and translated English are less lexically rich and use fewer idiomatic expressions and pronouns, and more explicit cohesive devices, than native English. Kruger and van Rooy (2016) similarly find non-native and translated English to be characterized by more explicit, formal choices and less lexical diversity than native English. Finally, Ivaska, Ferraresi, and Bernardini (2022) use the Universal Dependencies (UD) scheme to identify the syntactically defined bigrams of Parts-of-Speech (POS) that set translated and non-native English apart from non-translated, native English. Results

point to greater reliance on post-nominal noun phrase modification by means of prepositional phrases (pointing to explicitness) and a less heightened sensitivity to register conventions in the constrained varieties.

Kotze's framework (e.g., 2020, 2022) is the most thorough attempt to date to go beyond single comparisons and model varieties in terms of the different constraints operating on them. Her model includes a constraint matrix with five socio-cognitive dimensions (see further Section 3.2), conceived as continua: language activation (monolingual—bilingual), modality and register (spoken, written, multimodal), text production (independent—dependent), (language) proficiency (native/highly proficient user—learner), and task expertise (expert—non-expert) (Kotze 2022: 76-77). The interplay of these macro-level constraints and of less predictable micro-level ones (such as individual variation) is hypothesized to be reflected in the linguistic patterns of different constrained varieties, such that one could try to predict similarities and differences among varieties based on the constraints operating on them, and test these predictions empirically.

2.4 Learner translation corpora as a subset of learner corpora

Most of the studies reviewed in 2.3 have compared professional translation, L2 language production and native language production. However, obtaining large enough collections of such data produced by comparable populations in similar text production settings is nearly impossible, unless one resorts to quasi-artificial settings like the European Parliament (Ivaska et al., 2022). Given the well-known effects of even minor differences

in the datasets on the results obtained (cf. Callies, 2015; Hasselgård and Ebeling, 2018), such comparability issues should be carefully considered.

One way in which the impact of population- and setting-related issues on investigations of the constrained language hypothesis can be limited is by relying on translation and L2 learner data. Learner translation corpora (LTCs) are bilingual or multilingual databases including STs and multiple learner translations, aligned at the sentence level and often annotated with POS, lemma and error information. Granger and Lefer (2020:1184) define them as "two-in-one resources, as they contain translations produced by foreign language learners or trainee translators".

LTCs can be approached either from the parallel perspective (as source-target or target-target file pairs), or from the comparable perspective (as sets of TTs, to be compared to non-translated or "golden standard" translated texts). Interest in LTCs has recently seen a revival after the pioneering projects of the 1990's and 2000's (Bowker and Bennison, 2003; Castagnoli, Ciobanu, Kübler, Kunz, and Volanschi 2006). Several corpora have since seen the light, including the *UPF learner translation corpus* (English/Catalan; Espunya, 2014), the *Russian Learner Translator Corpus (*Kutuzov and Kunilovskaya, 2014), the *Undergraduate learner translator corpus* (Arabic; Alfuraih, 2020), and the *Multilingual Student Translation corpus* (MUST, Granger and Lefer, 2020).

Translation is taught in different ways and for different purposes: in translation departments, it is usually construed as a multidimensional competence, encompassing linguistic, cultural, interpersonal, technological and professional aspects, as well as interlinguistic transfer skills in the strict sense (EMT, 2017). In modern languages department, there has recently been a revival of "pedagogic translation" as an effective

instructional strategy "used to acquire linguistic, interlinguistic and intercultural competence in fields other than translation studies" (González-Davies, 2014: 8-9).

In terms of naturalness and authenticity, there are clear parallels between translational learner language and more familiar types of learner production: in the same way as "[a] dissertation written by a non-native speaker is clearly a part of normal academic business [and] writing an essay is a normal class activity, but is more restricted to the context of language learning" (Osborne 2015: 342), professionally- and pedagogically-oriented translations can also be set along a cline of naturalness. Also, if it makes sense to consider learner translation corpora as "peripheral learner corpora" (Gilquin, 2015: 10) when looked at from the perspective of language competence, they constitute core resources when investigating constrained language. As claimed by Granger and Lefer (2020: 1196),

> if used in combination with other corpora, the [LTC] allows for empirical investigations of student translation as compared with professional translation and/or learner/non-native writing, which can provide invaluable insights into interlingual mediation and constrained communication characterised by bilingual language activation.

3. Method

In this Section we present our research questions, describe our datasets, and detail the method used to extract, score and compare collocations across translations and essays, with a view to identifying the effects of TL salience, SL prominence, and connectivity.

3.1 Research questions and overall research method

In line with the constrained language hypothesis, our research questions address three variables which might affect translation and L2 writing differently as concerns their "collocationality", roughly defined as the presence of strongly associated lexical combinations. Drawing inspiration from Halverson's (2017) revised *gravitational pull* model, we distinguish three sources of constrainedness effects, all of which have been previously observed in both translation and learner language (Section 2). The first such source is *salience in the TL*, a term Halverson (2017: 13) uses to refer to the "idea that some patterns of activation within schematic networks will be more prominent than others, due to their higher frequency of use over time". The second source is *prominence in the SL*, reflecting the hypothesis that "highly salient representational elements in the source language […] would cause what is referred to as interference/transfer or cross-linguistic influence in second language acquisition research" (ibid.: 14). The third source is *connectivity*, i.e., the strength of the links between source and target equivalents, resulting from "high frequency co-occurrence of a translation pair" (ibid.). Connectivity is normally addressed as part of transfer effects in the learner language literature (cf. Gilquin, 2008). Yet it deserves special consideration in a study contrasting independent and dependent language production, since "[i]t is likely that the effect of connectivity in

an overtly bilingual task such as translation will be different from the effect in a more monolingual production mode" (Halverson 2017: 37).

Our research questions can be stated as follows:

- RQ1: Do translations into English and essays written in English by Italian translation students differ in levels of collocationality, as quantified by two association measures (AMs) that highlight salient vs. frequent combinations? [TL effects]
- RQ2: Are collocations produced in the two modes more, less or equally likely to result from the existence of direct cross-linguistic links between English and Italian? [connectivity effects]
- RQ3: Are Italian equivalents of English collocations used in translation overall more, less or equally collocational compared to the Italian equivalents of the collocations used in L2 writing? [SL effects]

The first part of the analysis (RQ1) is more quantitative in nature and concerns the status of collocations in the L2/TL: by relying on several AMs, we assess whether one text production mode is associated with higher or lower collocationality levels than the other; we operationalise collocationality as the median AM score of word pairs produced by students in either text production mode. In the second, more exploratory part (RQ2 and RQ3), we focus on the possible relationships between the collocations observed in English and their equivalents in Italian. Specifically, we narrow down the analysis to a manageable subset of highly frequent, non-idiosyncratic collocations and identify their equivalents using the translation STs, bilingual dictionaries and machine translation. The aim is to explore whether the differences observed in the two modes can be related to

direct cross-linguistic links (through lexicographic evidence) and equivalent collocations (through L1 frequency data).

3.2 Description of the dataset

Given the novelty of the comparison, in the present study we opted for an opportunistic data collection procedure. An enormous amount of effort is needed to build well-designed and balanced learner corpora, which doubles up in the case of different task types (and triples in the case of translation data, for which STs must be collected as well). We therefore believe it makes sense to evaluate advantages and disadvantages of this research framework before embarking in time-consuming corpus collection processes.

The varieties we compare are student essay writing in English (specifically term papers in corpus linguistics) and translation into English. Table 1 shows the constraint dimensions for these two varieties, based on Kotze's (2020) constraint model. In both cases language activation is bilingual, with the same directionality (L1 influences L2) and typological relation (Italian/English), since all students are native speakers of Italian who produce their texts in English. Modality and register are comparable across subcorpora (written, formal), but genres and topics differ: the L2 writing subcorpus contains essays in corpus linguistics, while the translation corpus contains argumentative/expositive texts on a variety of topics (economics, marketing, psychology). Text production, the main variable at stake, is unmediated for L2 writing and mediated for translation, in the sense that in the latter, but not in the former, "a prior text delimits and shapes production" (Kotze 2022: 346). Language proficiency and task expertise are constant, since all students belong to the same population. English proficiency is tested upon enrolment: the

threshold level is level C1 of the *Common European Framework of Reference for Languages* (Council of Europe 2020), as tested by the online Oxford Placement Test.

All students are enrolled in the international Master's degree in Specialized Translation at the University of Bologna, and have agreed to have their work included in this dataset and shared with the research community. The essays were produced as end-of-course exams for a module on corpus linguistics while the translations were produced either as homework or under exam conditions. In both cases students had access to reference materials and supports (such as computer-aided translation tools and specialised corpora).

Table 1. Constraints applying to the studied varieties (based on Kotze 2020)

	Variety: L2 writing	Variety: translation into the L2
Language activation	Bilingual	Bilingual
Modality and register	Written, formal (academic)	Written, formal (various genres)
Text production	Unmediated	Mediated
Proficiency	Proficient	Proficient
Task expertise	Semi-expert	Semi-expert

The essay corpus consists of 106 essays, by as many students, on 13 corpus linguistics topics, for a total of 224,968 words. The translation corpus consists of 131 translations of 37 STs, for a total of 27,393 (translated) words; the 19 contributing students provided from a minimum of two to a maximum of 12 translations each. No student contributed to both datasets. As can be observed in Table 2, the two subcorpora

differ substantially in several ways, including number of students, size, average text length and associated amount of variation (standard deviation; SD).

Table 2. Basic information about the dataset

	Essays	Translations
Number of students	106	19
Number of texts	106	131
Number of words	224,968	27,393
Average number of words per student	2,122 (SD: 444)	1,442 (SD: 1193)
Average number of words per text	2,122 (SD: 444)	209 (SD: 267)

All texts were anonymised and spot-checked for conversion and other mistakes. The essays were cleaned of parts that might skew the analyses (reference sections, prompts). All files were then converted from their original format (PDF/Ms Word) to text format, tagged, lemmatised, and parsed with the spaCy+UDPipe Universal Dependency Parser (Straka 2018; MIT TakeLab[1]).The datasets compared thus provide access to naturalistic learner production, representative of learner performance in the two tasks, but this comes at the cost of limited comparability. We address the ways in which our method tries to limit the impact of comparability issues in Section 3.3.2, and return to this point in a more programmatic way in the final Section (5).

3.3 Quantitative and qualitative analysis

3.3.1 Definition, extraction and scoring of TL/L2 collocations

The definition of collocation adopted in this study can be ascribed to the so-called "frequency approach", which sees collocations as groups of words displaying a statistical

[1] https://github.com/TakeLab/spacy-udpipe

tendency to occur near each other in texts (cf. Durrant (2014) for an overview of other approaches). Within the frequency approach, several operationalisations of collocation have been proposed varying according to parameters such as the nature of the linguistic items involved, the length of the sequence, the corpus search strategy and the statistics used to quantify co-occurrence (cf. Gries, 2008).

The last two parameters are particularly important. In the learner language and translation studies literature, corpus-based investigations of collocations[2] have typically selected bigrams based on the POS of their constituent words, e.g., adjectives and nouns (Siyanova and Schmitt, 2008), verbs and nouns (Laufer and Waldman, 2011), and adverbs and adjectives (Granger, 1998). Such bigrams are either searched for by selecting a node and scanning concordances or collocation lists for co-occurring words that belong to the preselected POS within a given span (e.g., Durrant, 2014), or by targeting shallow POS patterns in tagged corpora (e.g., Granger and Bestgen, 2014). To overcome problems associated with these approaches, which tend to exclude more complex structures and longer-distance relations, while potentially including badly formed pairs, in this study we rely on syntactic dependencies to extract collocation candidates. This allows us to disregard distance between the words in the pair, their constituent order and their hierarchy.

To quantify strength of attraction between collocates we rely on an AM which has been widely used in the literature, i.e., MI, and one that has been proposed more recently, i.e., logDice (Rychlý 2008). These measures have been selected for their ability to shed light on the preference for salient vs. frequent word combinations, since research on

[2] We do not include in this category lexical bundles, i.e., fixed, uninterrupted sequences of words typically of length above 2 (e.g., Paquot 2014).

learner language has shown that "learners are commonly found to be more proficient users of frequent collocations" than of "less frequent but more specialised […] collocations" (Ebeling and Hasselgård, 2015: 211; see also Ferraresi and Miličević, 2017 for similar results concerning translation). Specifically, MI is known to emphasise low frequency, strongly associated word pairs (ibid.), such as, in our corpus, *pivotal importance*, or *career path*. By contrast, logDice "highlights exclusive but not necessarily rare combinations" (Gablasova, Brezina, and McEnery, 2017: 10), such as *pay attention*, and *improve quality*.

A third AM (t-score) was also used as a preliminary filter to make sure that none of the analysed collocations would be extremely rare and/or specialised, for reasons that will be made clear in Section 3.3.2.

The collocation extraction procedure started with the selection of dependency relationships (see Figure 1 for an example of the output of the dependency parser).[3] The selection aimed at excluding grammatical co-occurrence phenomena and reproducing the patterns investigated in previous work (e.g., noun pre-modification by adjectives).

[3] The full UD tagset is available at https://universaldependencies.org.

Many	ADJ	many	amod	case	NOUN
cases	NOUN	case	nsubj	show	VERB
show	VERB	show	ROOT	show	VERB
,	PUNCT	,	punct	show	VERB
in	ADP	in	case	position	NOUN
this	DET	this	det	position	NOUN
position	NOUN	position	obl	show	VERB
,	PUNCT	,	punct	show	VERB
a	DET	a	det	preference	NOUN
positive	ADJ	positive	amod	preference	NOUN
semantic	ADJ	semantic	amod	preference	NOUN
preference	NOUN	preference	obj	show	VERB

Figure 1. An example of the parser output

Each token in the corpus is placed on a separate line (e.g., *cases*), followed by its POS (*NOUN*), lemma (*case*), syntactic dependency (*nsubj*), syntactic head (*show*) and POS of the head (*VERB*).

We selected 16 relationships involving adjectives, nouns, verbs, or adverbs, and focused on word lemmas rather than word forms, to limit data sparseness problems. Table 3 displays examples of lemma pairs representing the top 10 relationships in frequency order (combining the two corpora).[4]

[4] The remaining relationships are: *csubj* (causal subject), *fixed* (multiword expression), *nmod:npmod* (noun phrase as adverbial modifier), *iobj* (indirect object), *nmod:tmod* (temporal modifier), *compound:prt* (phrasal verb particle).

Table 3. Examples of extracted lemma pairs and dependencies

Relationship label (meaning)	Extracted triplet	Collocation
amod (adjectival modifier)	key_amod_element	key element
nmod (nominal modifier)	time_nmod_period	time period
advmod (adverbial modifier)	completely_advmod_different	completely different
conj (conjunct)	economic_conj_social	economic [and] social
obj (object)	skill_obj_acquire	acquire [new] skills
compound (compound)	press_compound_release	press release
nsubj (nominal subject)	study_nsubj_show	[the] study shows
acl (clausal modifier of noun)	create_acl_ability	ability [to] create
acl:relcl (relative clause modifier)	glue_acl:relcl_element	elements [which] glue [them together]
xcomp (open clausal complement)	available_xcomp_make	make [them more easily] available

The next step consisted in assigning AM scores to the extracted pairs. Following a procedure that, starting with Durrant and Schmitt (2009), has been used in several works in both learner corpus research and translation studies (e.g., Granger and Bestgen, 2014; Bernardini, 2011), frequency data for the calculation of AMs were derived from a large reference corpus of English, as a way of approximating the strength of association in general English.

The reference corpus used for this purpose is a 130-million-word dependency-parsed subset of the ukWaC corpus (Baroni et al., 2009).[5] The adoption of the same dependency scheme for the study corpus and the reference corpus made it possible to calculate frequencies of co-occurrence of lemmas within a given syntactic relationship, and then match them across the two corpora. For example, frequencies for the collocation candidate *tool_nmod_use* ("[the] use [of] tools") were obtained separately from those of

[5] The relatively small size of the subset is justified by the fact that dependency parsing is computationally costly. The spacy+UDPipe parser can process files of less than 1 MB on a 2016 MacBook Pro, which made it necessary to split the ukWaC subset into more than 600 files.

tool_obj_use ("[to] use [a] tool"). T-score and MI values were calculated using an ad hoc Python wrapper for the NLTK Collocations library;[6] the logDice calculation was implemented in the wrapper following the formula by Rychlý (2008).

3.3.2 By-subject analysis of collocationality in translations and essays

While in many ways inevitable, the factors that differentiate the learner essays and translation corpus (cf. Section 3.2) cannot be disregarded. In the quantitative comparison of collocationality in translations and essays we had to strike a balance between minimising confounding variables and not deliberately hiding differences between the datasets.

One of the potential distorting factors in the comparison is the degree of technicality of the texts associated with their topic and genre (cf. Paquot, 2014). Since translations may contain a higher variety of technical or scientific terms compared to essays, word pairs were filtered in two ways. First, we only retained pairs in which both lemmas are present in both translations and essays, regardless of the relationship, so that highly idiosyncratic pairs are excluded. Second, the analysis was restricted to candidate pairs with a ukWaC-derived t-score value of at least 10 and a frequency of at least 5 (cf. Durrant and Schmitt, 2009: 170); such threshold is expected to reduce the number of specialised terms.

Moreover, several students contributed more than one text to the translation dataset, which might violate the assumption of independence of data points.[7] To overcome the

[6] https://www.nltk.org/howto/collocations.html.

[7] This point was made by an anonymous reviewer. In a study in which subjects contributing texts are known, this problem can be avoided by adopting alternative analytical procedures. However, the observation might call for wider methodological reflection on the reliability of several statistical tests – including mixed models using texts rather than subjects as random factors – when the subject (e.g. the author/translator) is not known, as is often the case in corpus-based studies.

problem, we calculated the median values of MI and logDice of pairs matching the above-mentioned thresholds produced by each student, aggregating data at the level of subjects rather than texts, and thus obtaining independent observations across the two datasets.

The procedure still left us with an imbalanced number of observations, with the translation dataset also being very small in absolute terms (19 observations vs. 106 for the essays). For this reason, we opted for a statistically robust procedure known as bootstrapping, which consists in "randomly resampl[ing] from an observed data set to produce a simulated but more stable and statistically accurate outcome" (Plonsky et al. 2015); bootstrapping has been claimed to yield reliable results with samples as small as n=10 (Larson-Hall and Herrington, 2010: 381). Specifically, we carried out bootstrapped independent sample t-tests with 10,000 replications using the BCa method (cf. LaFlair et al. 2015: 51), comparing AM values by subject across translations and essays. The datasets were tested for normality and homogeneity of variance using Shapiro–Wilk and Levene's tests. Following LaFlair et al. (2015), the results of the bootstrapping procedures were inspected by checking skewness and kurtosis of the bootstrap values, normality of Q-Q plots, as well as outliers in jackknife-after-boot plots. As a further check that the imbalance between samples would not impact on results, all tests were repeated on three different samples in which we randomly downsized the essay dataset to the same size as the translation one (n = 19).[8] Effect sizes for the tests (*r* values) were calculated using the formula by Field et al. (2012: 384-385); in this case too, bootstrapping was performed on 10,000 samples to obtain more reliable statistics (along the lines of LaFlair et al. 2015: 72-73).[9]

[8] The sanity checks confirmed the results obtained in the main analyses; results are not reported in Section 4 below.
[9] All tests were performed in R (R Core Team 2022), using the *boot* package (Canty and Ripley, 2021) for bootstrapping.

3.3.3 Qualitative exploration of connectivity: data from bilingual dictionaries

We used lexicographic information to identify connectivity across translation and independent writing: a translation pair present in at least one bilingual dictionary is hypothesised to be both strongly connected and salient (i.e., worth pointing out to dictionary users). This is admittedly a rough operationalisation of connectivity, which differs from other approaches proposed in the literature (Halverson, 2017; Gilquin, 2008), and that will therefore require further reflection and refining in future work.

Due to the manual perusal required, only a limited portion of the full collocation dataset could be annotated with lexicographic information in this way. To make it possible to explore both similarities and differences across the varieties considered, collocations were grouped into three subsets:

- shared collocations: present in the same form in translations and essays, produced by at least two students in the translation dataset, and at least three students in the essay dataset;
- pairs unique to translations: present only in the translation dataset, produced by at least three students;
- pairs unique to essays: present only in the essay dataset, produced by more than six students.

The same constraints described in Section 3.2 apply here: collocations had to a) be formed of lemmas which are present in both corpora, and b) have t-score > 10 and fq ≥ 5 in ukWaC. To exclude idiosyncratic pairs from the analysis, the number of students, rather than the number of texts, was used as a criterion since in the translation dataset every student provided more than one text. The minimum number of students was set arbitrarily,

so as to obtain subsets of collocations roughly proportional to the sizes of the original datasets.

We checked each lemma pair in two bilingual Italian-English dictionaries from different lexicographic traditions, i.e., the Cambridge English-Italian Dictionary,[10] and the Zanichelli Ragazzini dictionary.[11] For each pair we recorded presence in at least one dictionary as a headword or an example. Several pairs were excluded in this phase, either because a dictionary flagged them as specialised (e.g. *limited_amod_company*), or because they were part of sequences longer than two words (e.g. *make_compound_process* and *decision_compound_process*, both part of the trigram *decision-making process*). Table 4 shows the number of collocations retained in each subset, with an indication of the three most frequent syntactic relationships in which they enter.

Table 4. Collocations and their syntactic relationships

Collocation subset	Number of collocations	Most frequent relationships (top 3)
Shared	34	*amod*, *advmod*, *obj*
Essays only	103	*amod*, *obj*, *advmod*
Translation only	54	*amod*, *obj*, *conj*
TOTAL	191	

Due to the limited number of observations and the tentative nature of the method, for this part of the analysis we did not carry out any statistical test.

[10] https://dictionary.cambridge.org/dictionary/italian-english/
[11] https://www.zanichelli.it/ricerca/prodotti/il-ragazzini-2021

3.3.4 Exploration of SL effects: equivalents from dictionaries, machine translation and source texts, with association scores from an SL corpus

Drawing on the same subset of collocations used to explore connectivity, we conducted a second exploration focusing on the third pole of Halverson's (2017) model, i.e., L1/SL effects on L2/TL production. By only focusing on frequent English combinations, the effects we observe, if any, are related to positive transfer, i.e., cases in which the influence of the L1/SL leads to correct rather than erroneous usage (Gilquin 2008: 5).

The first problem we encountered consisted in identifying the most plausible Italian equivalent for the collocations observed in learner production. Gilquin (2008: 14) mentions the possibility of adopting reversed translation, which relies "on the researcher's capacity to translate learners' interlanguage back into their mother tongue", but which therefore stands as an exception to the more objective methods of corpus-based research. To make the search as independent as possible from intuition, we relied on three sources of evidence:

- whenever a collocation was found in a bilingual dictionary, equivalents provided there were recorded;
- if no equivalents were found in dictionaries, Google Translate[12] was used to translate the collocation. Co-text derived from learner texts was used as input to obtain more pertinent results;
- for the translation and shared pairs subset, Italian STs were also scrutinised. If no equivalent expression was found, e.g., because of translation shifts between ST and TT (e.g. *more likely* as a translation of *tendono* [they tend]) or if the ST expression was not a bigram (e.g. *better quality* as a translation of *miglioramento* [improvement]), we marked the equivalent as 'not available'.

[12] https://translate.google.com.

The second issue we were faced with had to do with deciding which equivalent to test for SL frequency/salience when several options were available. ST equivalents from translations were not prioritised, so as not to insert a bias in the analysis. Instead, we calculated AMs for all SL equivalents and selected the one with the highest MI or logDice value.

In practice, each 2-word Italian equivalent of an English collocation was first manually lemmatised, and then looked up in a syntactically parsed 140-million-word subset of the itWaC corpus, a large reference corpus of Italian comparable to ukWaC (Section 3.3.1). Thanks to the cross-linguistic comparability of the UD annotation scheme, AM scores were calculated for combinations of lemmas entering the same syntactic dependencies as for English.[13] This made it possible to assign MI and logDice scores to all equivalents in the dataset; if the observed frequency was lower than 5, however, equivalents were discarded, as the corresponding AM scores might be unreliable. Equivalents were ranked in decreasing order of MI and logDice scores, and the one with the highest score for each AM was retained.

Table 5 summarises the research questions, and the data and methods used to address them.

Table 5. Summary table of RQs, data and methods used in the study

Research question	Data	Method
RQ1: TL effects	Median MI and logDice scores of word pairs by student in translations and essays. Word pairs = syntactically defined lemma bigrams, where: - both lemmas are present in translations and essays; - fq ≥ 5 and t-score ≥ 10 (ukWaC)	Bootstrapped t-tests comparing median MI and logDice scores

[13] Since we could not anticipate all the possible relationships of the two lemmas in Italian, frequencies of all lemma combinations were summed up.

RQ2: connectivity effects	Italian (SL) equivalents of a subset of the word pairs as defined for RQ1: • shared collocations: by at least 2 students in translations, at least 3 students in essays; • pairs unique to translations: by at least 3 students; • pairs unique to essays: by more than 6 students. Equivalents from 2 bilingual Italian-English dictionaries	Comparison of the number of SL-TL equivalences included vs. not included in bilingual dictionaries
RQ3: SL effects	MI and logDice scores of SL equivalents of word pairs from RQ2. Equivalents from: - bilingual dictionaries (see RQ2); - Google Translate (if not in dictionaries); - STs, where available. Only SL equivalents with fq ≥ 5 (itWaC) were retained. When multiple equivalents were present, the one with the highest MI and highest logDice score was retained	Comparison of median MI and logDice values of English pairs and of Italian equivalents

4. Results

Translations display higher levels of collocationality than essays, both in terms of MI and logDice. More specifically, the by-subject median MI values of syntactically defined lemma pairs are significantly higher in translations (M = 5.12, SD = 0.37) than essays (M = 4.34, SD = 0.35). This is shown by the bootstrapped t-test (t = –8.85; bias = –0.04), where the 95% BCa confidence interval not including 0 [–11.826, –6.356] points to statistical significance (Plonsky et al. 2015: 600). The bootstrapped effect size is large (*r* = 0.62; bias = –0.004), with a 95% confidence interval of [0.49, 0.73].

A significant difference is also observed in terms of logDice values (M_{trans} = 7.07, SD_{trans} = 0.40; M_{essay} = 6.80, IQR_{essay} = 0.30), with boostrapped t = –3.45 (bias = –0.02) and a BCa confidence interval of [-6.01, -0.98]; in this case the effect size is weaker (*r* = 0.30; bias = –0.002), and the confidence interval very large [0.09, 0.48], indicating more

uncertainty about this particular result (in practice, the effect size for this comparison may vary between 0.09 and 0.48). The results are plotted in Figure 2.

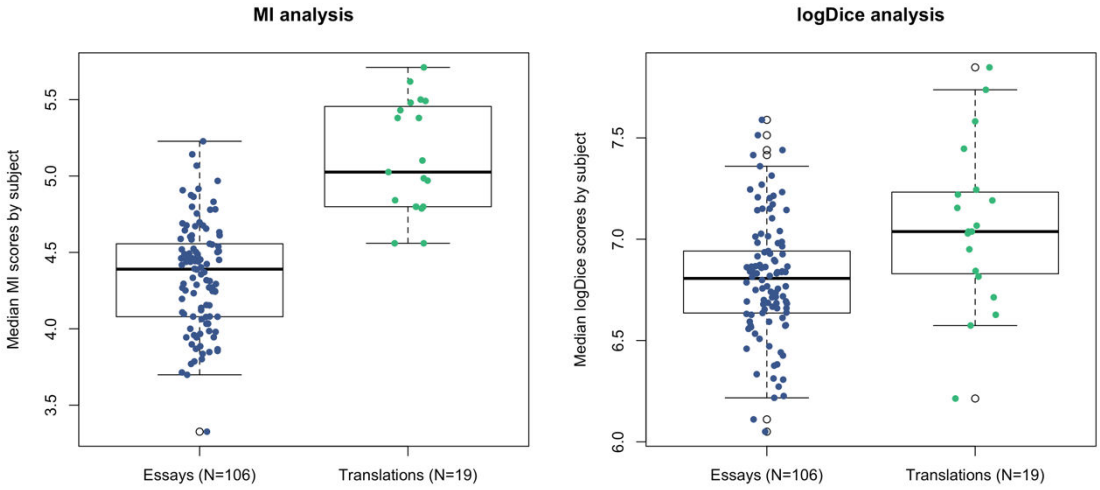

Figure 2. By-subject analysis of collocationality in essays and translations

Moving on to our exploratory analysis, we first investigated the level of connectivity between a subset of collocations and their SL equivalents, based on information from bilingual dictionaries (Section 3.3.1). The mosaic plot in Figure 3 shows the proportion of pairs a) only observed in essays, b) only observed in translations, and c) shared by both modes, which are attested in at least one bilingual dictionary as either a head word or part of an example. The proportion of collocations of lexicographic import observed in essays and translations is roughly the same: about one third of the respective total. This proportion is substantially higher in the case of shared pairs: slightly over half of the total. To get a rough idea of the pairs present in/absent from the dictionaries, Table 5 shows selected examples from the three subsets, representing the adjectival modifier dependency relation (*amod*).

Table 5. English collocations from the three subsets found/not found in at least one bilingual dictionary

	Present in at least one dictionary	Absent from dictionaries
Translations	*social environment, open space, strong link*	*several decades, national level, independent body*
Essays	*distinctive feature, important role, better understanding*	*large range, second part, different pattern*
Shared	*great importance, starting point, key role*	*first one, significant difference, key element*

If one accepts explicit mention in a bilingual dictionary as an indication of cross-linguistic salience/connectivity, these results would suggest that the cross-linguistic salience/connectivity of a collocation pair increases the likelihood of its TL component being used by learners, regardless of the type of constrained text production they are engaging in.

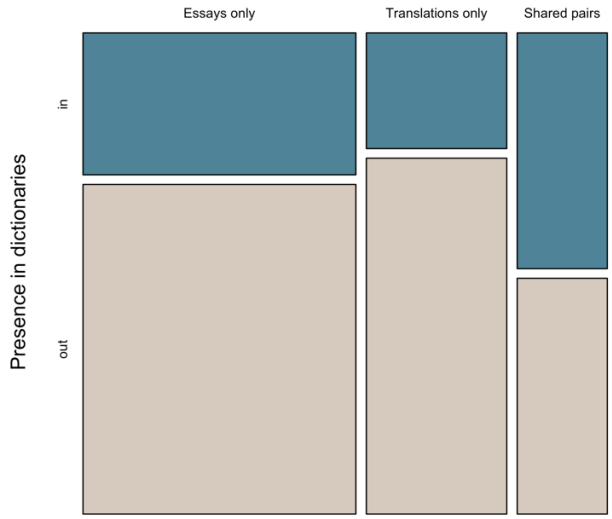

Figure 3. Dictionary evidence about cross-linguistic connectivity of collocation pairs

Finally, we checked for SL effects by comparing the same three subsets (translation-only, essay-only and shared) in terms of the relation between the median MI and logDice values of the selected English pairs and the corresponding values of their Italian equivalents. As shown in Table 6, median values are consistently lower for L2/TL collocations than for their actual or hypothesised SL/L1 equivalents, with a single exception: the logDice values of essay-only pairs are higher in English than Italian (6.21 vs. 6.77). For the shared and translation-only pairs, TL median logDice values are lower. Here we may be observing a difference related to the mode of constrained language production: the collocations found in the students' translations, and those shared by translations and essays, would appear to derive from *even more* frequent/standard collocations in the SL. This is not the case for essay-only collocations, whose Italian translation equivalents are *less* frequent/standard in Italian than in English.

The MI comparison returns higher values for SL than TL collocations across all three subsets. The largest difference is observed for translation-only pairs, whose median MI value is 6.30 in the SL, and 4.87 in the TL. Regardless of their mode of production, collocations found in the English texts seem to derive from more salient collocations in Italian. While some of the absolute values of SL equivalents might be overestimated, due to our choice of using the highest collocational value in case of multiple equivalents (see Section 3.3.4), this feature should apply equally across subsets.

Table 6. MI and logDice comparisons for the selected essay-only, translation-only, and shared pairs

	logDice		MI	
	SL/L1 (Italian)	TL/L2 (English)	SL/L1 (Italian)	TL/L2 (English)
Translations	6.78	6.57 (-)	6.30	4.87 (-)
Essays	6.21	6.77 (+)	4.40	4.32 (-)
Shared	8.37	8.20 (-)	6.55	5.97 (-)

5. Discussion and conclusion

In this paper we have investigated collocations in learner language by comparing L2 writing and translation, two constrained language varieties differing in terms of type of text production (unmediated vs. mediated). The approach was exploratory and aimed to first of all ascertain that a common ground could be found on the basis of which to compare the collocations found in two small, opportunistically assembled datasets representing a very homogeneous learner population (students from a single Master's course).

Based on Halverson's (2017) revised gravitational pull hypothesis and Kotze's (2022) constrained communication framework, our research questions addressed three types of influence observed in bilingual mediated and unmediated language production: TL frequency/salience, SL-TL connectivity, and SL prominence. On the basis of a comparison of median values of MI and logDice across subjects in the two subcorpora, we have answered our RQ1 (are translations more/less collocational than essays?) by showing that translations are significantly more collocational than essays both in terms of frequent combinations (higher logDice values) and salient combinations (higher MI values).

To answer RQ2 and RQ3 we relied on an exploratory method, manually identifying candidate SL equivalents and checking their presence in bilingual dictionaries (as a way of measuring cross-linguistic connectivity, RQ2) and their collocational status in the SL (as a way of measuring SL prominence, RQ3). Connectivity was found to be higher for lemma combinations shared by both modes of production, while no difference was observed for translation- and essay-specific combinations (RQ2). In other words, the existence of a cross-linguistic link would not seem to affect choice of a collocation in translation vs. essay writing differently. If a lemma pair appears in both translations and essays, however, it is more likely that it is cross-linguistically connected (or at least lexicographically interesting), than if it were used in either mode only.[14]

Lastly, SL influence was observed in both varieties, since the median values of AMs in the TL correspond quite closely to those in the SL for all subsets (RQ3). In the case of essays, the median logDice values are higher in the TL than the SL, suggesting that these learners, who were taught and had read about the topic(s) of their essays for one semester, had probably already interiorised the common collocations in the field to a greater extent than the more salient, but rarer, high-MI collocations. The result concerning translation is even more revealing. While the median values for MI in translation observed in our quantitative analysis are significantly higher than in essays, this might be a carry-over effect from corresponding higher median values in the SL. This should not downplay the importance of the fact that students are able to produce salient TL collocations matching salient SL collocations, even when constrained by the need to convey "the form and sense of the original as accurately as possible" (Rendall, 1997: 155).

[14] While one might hypothesise that this finding is simply due to the fact that the learners looked up these equivalences in the bilingual dictionaries themselves, this seems unlikely given the rather general, unspecialised nature of the collocations (Table 5).

While most studies compare learner and/or translated language to native/non translated language, in this study we have kept the bilingual constraint constant, since both tasks involve the activation of bilingual competences, and instead attempted to isolate the mediation constraint, which distinguishes translation from L2 learner writing. Out of Halverson's three sources of effects, TL salience is indeed confirmed to be more specifically translational, affecting mediated language production more strongly than unmediated language production. Connectivity and source language prominence are instead found to play an important role in both modes of bilingual communication, but not to affect the two modes differently. Based on these first results, and keeping in mind the many methodological limitations that we have been faced with – to which we will come back shortly – we would point to TL salience as a more specifically translational effect, and to connectivity and SL influence as more likely candidates for general bilingual processing effects, affecting both translation and second language use similarly.

As a key methodological point, we would like to further reflect on the challenges involved in comparing different types of naturalistic learner data. The typical translation tasks that students in translation degrees are likely to be confronted with concern text types and domains that inevitably differ with respect to those that they are asked to produce as writing tasks. In such a scenario, the authenticity of the datasets is arguably impossible to reconcile with their comparability. To tackle the comparability issue without compromising on authenticity, our choice was to address confounding variables at the data analysis stage rather than corpus building stage. Thus, we only focused on a specific subset of non-technical collocations to reduce the potential bias introduced by differences in the level of specialisation of texts, and we adopted a statistical procedure

in the quantitative analysis that takes into account differences in corpus designs and corpus size.

Despite these methodological hurdles, we would claim that the comparison between learner translation and learner writing is one worth undertaking. Since translation is interpretive language use (Gutt, 1991), some of the cognitive effort needed for idea generation, planning etc. in independent writing is freed, and can be employed for more careful monitoring of linguistic choices (Lanstyak and Heltái, 2012). While Lanstyak and Heltái (ibid.) consider the monitor ability as a distinguishing feature of translators when compared to other bilinguals, it might equally be construed as an effect of the task. This has implications for learner corpus research in general: relying solely, or mainly, on essay writing, a highly complex task especially when carried out in an L2, might lead us to underestimate L2 learners' collocational competence. Our research design, which compares the performance of bilingual students drawn from the same population but engaging in independent vs. dependent text production, has allowed us to appreciate differences in their ability to produce frequent and salient collocations in response to different task demands. This confirms Granger's (2015: 12) view that "research would benefit from more studies focused on variables other than L1 transfer, such as the effect of foreign vs. second language setting, task effects or differences in proficiency level".

A further advantage of using learner translation data in investigations of learner language is the availability of SL equivalents. The problems inherent in the identification of such equivalents have been pointed out in the learner corpus literature (notably by Gilquin, 2008). Naturalistic translation data provide a means of evaluating the subjective distance between SL and TL elements, since the ST contains the expression that the student reacted to when generating a given TT expression. Importantly for corpus

research, cross-linguistic equivalence is not established on the basis of intuition or elicited data, but through an instance of authentic, communicative language use (though of a dependent rather than independent kind).

Now that learner translation corpora collected through explicit criteria and annotated with rich, standardised metadata are becoming more widely available for a large number of language pairs (e.g. through the MUST project; Granger and Lefer, 2020), it should be possible to compare more representative and balanced datasets than was the case in this work, thus attaining a better understanding of the different constrained language varieties produced by the growing community of multicompetent language users across the globe.

References

Alfuraih, R. F. (2020). The undergraduate learner translator corpus: a new resource for translation studies and computational linguistics. *Language Resources and Evaluation, 54*, 801–830.

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics, 22*(2), 173–195.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources & Evaluation, 43*, 209–226.

Bernardini, S. (2011). Monolingual comparable corpora and parallel corpora in the search for features of translated language. *SYNAPS, 26*, 2–13.

Bowker, L., & Bennison, P. (2003). Student translation archive: Design, development and application. In F. Zanettin, S. Bernardini & D. Stewart (Eds.), *Corpora in translator education* (pp. 103–117). London and New York: Routledge.

Callies, M. (2015). Learner corpus methodology. In S. Granger, G. Gilquin, & F. Meunier, (Eds.) *The Cambridge handbook of learner corpus research* (pp. 35–55). Cambridge: Cambridge University Press.

Canty A., & Ripley B.D. (2021). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-28.

Castagnoli, S., Ciobanu, D., Kübler, N., Kunz, K., & Volanschi, A. (2006) Designing a learner translator corpus for training purposes. *Proceedings of TALC2006*, 1–19.

Council of Europe (2020) *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Strasbourg: Council of Europe Publishing.

Crossley, S. A., Salsbury, T., & Mcnamara, D. S. (2015) Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics, 36*(5), 570–590.

Dayrell, C. (2007). A quantitative approach to compare collocational patterns in translated and non-translated texts. *International Journal of Corpus Linguistics, 12*(3), 375–414.

Durrant, P. (2014). Corpus frequency and second language learners' knowledge of collocations: A meta-analysis. *International Journal of Corpus Linguistics, 19*(4), 443–477.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics, 47*(2), 157–177.

Durrant, P., & Siyanova-Chanturia, A. (2015). Learner corpora and psycholinguistics. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 57–78). Cambridge: Cambridge University Press.

Ebeling, S., & Hasselgård, H. (2015). Learner corpora and phraseology. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 207–230). Cambridge: Cambridge University Press.

Ellis, N. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics, 32*, 17–44.

Ellis, N. C., Simpson-Vlach, R. & Maynard, C. (2008). Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 41*(3), 375–396.

EMT (European Master's in Translation) (2017). Competence Framework 2017. Retrieved from https://ec.europa.eu/info/sites/default/files/emt_competence_fwk_2017_en_web.pdf

Espunya, A. (2014). The UPF learner translation corpus as a resource for translator training. *Language Resources and Evaluation, 48*, 33–43.

Feng, H. (2020). *Form, meaning and function in collocation. A corpus study on commercial Chinese-to-English translation*. London and New York: Routledge.

Ferraresi, A. and Miličević. M. (2017) Phraseological patterns in interpreting and translation. Similar or different? In G. De Sutter, M.-A. Lefer, & I. Delaere

(Eds.), *Empirical translation studies: New theoretical and methodological traditions* (pp. 157–182). Berlin: Walter de Gruyter.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London: Sage Publishing.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning, 67*(S1), 155–179

Gilquin, G. (2008). Combining contrastive and interlanguage analysis to apprehend transfer: detection, explanation, evaluation. In G. Gilquin, S. Papp, & B. Díez-Bedmar (Eds.), *Linking contrastive and learner corpus research* (pp. 3–33). Amsterdam and New York: Rodopi.

Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin, & F. Meunier, (Eds.) *The Cambridge handbook of learner corpus research* (pp. 9–34). Cambridge: Cambridge University Press.

González-Davies, M. (2014). Towards a plurilingual development paradigm: From spontaneous to informed use of translation in additional language learning. *The Interpreter and Translator Trainer, 8*(1), 1–14.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 145–160). Oxford: Clarendon Press.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics, 52*(3), 229–252.

Granger, S., & Lefer, M.-A. (2020). The *Multilingual Student Translation* corpus: a resource for translation teaching and research. *Language Resources and Evaluation, 54*, 1183–1199.

Gries, S. T. (2008). Phraseology and linguistic theory. A brief survey. In S. Granger, & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 3–26). Amsterdam and Philadelphia: John Benjamins.

Grosjean, F. (2013). Speech perception and comprehension. In F. Grosjean, & P. Li (Eds.), *The psycholinguistics of bilingualism*, (pp. 29–49). Oxford: Blackwell Publishers.

Halverson, S. L. (2017). Gravitational pull in translation. Testing a revised model. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical translation studies: New theoretical and methodological traditions* (pp. 9–46). Berlin: de Gruyter.

Hasselgren, A. (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), 237–260.

Hasselgård, H. (2019). Phraseological teddy bears. Frequent lexical bundles in academic writing by Norwegian learners and native speakers of English. In V. Wiegand & M. Mahlberg (Eds.), *Corpus linguistics, context and culture* (pp. 339–62). Berlin: De Gruyter.

Hasselgård, H., & Ebeling, S. (2018). At the interface between Contrastive Analysis and Learner Corpus Research: A parallel contrastive approach. *Nordic Journal of English Studies, 17*(2), 182–214.

Hoey, M. (2005). *Lexical priming. A new theory of words and language*. London and New York: Routledge.

Ivaska, I., Ferraresi, A., & Bernardini, S. (2022). Syntactic properties of constrained English: A corpus-driven approach. In S. Granger, & M.-A. Lefer (Eds.) *Extending the scope of corpus-based translation studies*. (pp. 133–157). London: Bloomsbury.

Jantunen, J. H. (2004). Untypical patterns in translations. In A. Mauranen, & P. Kujamäki (Eds.), *Translation universals: Do they exist?* (pp. 101–128). Amsterdam and Philadelphia: John Benjamins.

Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. New York and London: Routledge.

Kjellmer, G. (1984). Some thought on collocational distinctiveness. In J. Aarts & W. Meijs (Eds.), *Corpus linguistics*. *Recent Developments in the Use of Computer Corpora in English Language Research* (pp 163-71). Amsterdam: Rodopi.

Kellerman, E. (1977). Towards a characterization of the strategy of transfer in second language learning. *Interlanguage studies Bulletin, 2*, 58–145.

Kenny, D. (2001). *Lexis and creativity in translation. A corpus-based approach.* Manchester: St. Jerome.

Kolehmainen, L., Meriläinen, L., & Riionheimo, H. (2014). Interlingual reduction: Evidence from language contacts, translation and second language acquisition. In H. Paulasto, L. Meriläinen, H. Riionheimo, & M. Kok (Eds.), *Language contacts at the crossroads of disciplines* (pp. 3–32). Newcastle: Cambridge Scholars Publishing.

Kotze, H. 2020. Converging what and how to find out why. In L. Vandevoorde, J. Dams, & B. Defrancq, (Eds.), *New empirical perspectives on translation and interpreting* (pp. 333-71). Oxford: Routledge.

Kotze H. 2022. Translation as constrained communication: Principles, concepts and methods. In S. Granger & M.-A. Lefer (Eds.), *Extending the Scope of Corpus-based Translation Studies* (pp. 67-98). London: Bloomsbury.

Kruger, H., & van Rooy, B. (2016). Constrained language. A multidimensional analysis of translated English and a non-native indigenised variety of English. *English World-Wide, 37:1*, 26–57.

Kutuzov, A., & Kunilovskaya, M. (2014). Russian learner translator corpus. In P. Sojka, A. Horak, I. Kopecek, & K. Pala (Eds.), *Text, speech and dialogue* (pp. 315–323). Berlin: Springer.

LaFlair, G. T., Egbert, J, & Plonsky, L. (2015). A practical guide to bootstrapping descriptive statistics, correlations, t tests, and ANOVAs. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research*, (pp. 46–77). New York: Routledge.

Lanstyak, I., & Heltai, P. (2012). Universals in language contact and translation. *Across Languages and Cultures, 13*(1), 99–121.

Larson-Hall, J., & Herrington, R. (2010). Improving data analysis in Second Language Acquisition by utilizing modern developments in applied statistics. *Applied Linguistics*, *31*(3), 368–390.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*(2): 647–672.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam and Philadelphia: John Benjamins.

Osborne, J. (2015). Transfer and learner corpus research. In S. Granger, G. Gilquin, & F. Meunier, (Eds.) *The Cambridge handbook of learner corpus research* (pp. 333–356). Cambridge: Cambridge University Press.

Paquot, M. (2014). Cross-linguistic influence and formulaic language: Recurrent word sequences in French learner writing. In L. Roberts, I. Vedder, & J. Hulstijn (Eds.), *EUROSLA Yearbook* (pp. 216–237). Amsterdam and Philadelphia: John Benjamins.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards, & R. Schmidt (Eds.), *Language and Communication* (pp. 191–226). London: Longman.

Plonsky, L., Egbert, J., & LaFlair, G.T. (2015). Bootstrapping in applied linguistics: Assessing its potential using shared data. *Applied Linguistics, 36*(5), 591–610.

R Core Team. (2022). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Rabinovich, E., Nisioi, S., Ordan, N., & Wintner, S. (2016). On the similarities between native, non-native and translated texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1870–81.

Rendall, S. (1997). The translator's task, Walter Benjamin (Translation). *TTR: Traduction, Terminologie, Rédaction, 10*(2), 151–165.

Rychlý, P. (2008). A lexicographer-friendly association score. *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, 6–9.

Shlesinger, M. (1989). *Simultaneous interpretation as a factor in effecting shifts in the position of texts on the oral-literate continuum*. MA thesis, Tel Aviv University.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Siyanova, A., Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *The Canadian Modern Language Review, 64*(3), 429–458.

Wang, Y. (2016). *The idiom principle and L1 influence: A contrastive learner-corpus study of delexical verb + noun*. Amsterdam and Philadelphia: John Benjamins.

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.