Collocations in Contact: Exploring Constrained Varieties of English through Corpora

Adriano Ferraresi

Abstract
This contribution presents a study of collocations in a subset of the EPTIC corpus, which encompasses comparable discourse samples in English produced under a variety of communicative constraints (oral and written, original and mediated, native and non-native). Results are interpreted against the backdrop of Lanstyák and Heltai's (2012) "constrained language" hypothesis, according to which translations, on a par with other varieties of English produced in language contact settings, share linguistic features that set them apart from monolingual native production. Evidence of such constraints is observed in this study with respect to collocational patterns defined by Mutual Information values. Findings are discussed in the light of the cognitive constraints characterising the different language production tasks under investigation.

Keywords: constrained communication, non-native language, translation, interpreting, collocations.

1. Introduction: Translation, language contact, and the constrained language hypothesis

Translating a text into one's first language and delivering a speech in a foreign language share a fundamental trait: two languages are activated simultaneously during the language production task. According to a recent hypothesis by Lanstyák and Heltai (2012), this shared condition of bilingual activation might be associated with similar cognitive constraints, despite obvious differences in terms of tasks and individuals' language competence. Such shared constraints might in turn explain linguistic phenomena that have been observed independently in "traditional" contact varieties, such as second language use, and in translation. The authors (Lanstyák and Heltai 2012: 114) suggest that research in these fields has much to gain from developing a unified framework in which broadly conceived contact varieties are seen as related instances of "constrained" communication[1] (see also Kranich (2014: 96) for a similar argument on translation as a "locus of language contact").

Theoretical work in this area has singled out several phenomena that have been hypothesised to be typical of single varieties, but which in fact might be common to constrained varieties at large. Kolehmainen, Meriläinen and Riionheimo (2014), for instance, review literature on interlingual reduction in second language acquisition and translation studies, arguing that similar linguistic features "have been examined within different theoretical frameworks and with differing terminology" (p. 4). "Avoidance", "underproduction" and "simplification" observed in second language use (Jarvis and Pavlenko 2008) are suggested to be akin to translators' tendency to underuse "unique items" (Tirkkonen-Condit 2004), i.e. items in the target language which have no straightforward equivalents in the source language.

Empirical studies exploring similarities and differences between constrained varieties bottom-up are rarer. Work by Rabinovich, Nisioi, Ordan and Wintner (2016) and Kruger and van Rooy (2016) constitute notable exceptions. In both studies, a corpus-based comparison is carried out between English translations, non-native and native written texts, with a view to identifying shared linguistic traits in the two constrained varieties that set them apart from the unconstrained one. Findings consistently suggest that translations and non-native texts display a higher degree of formality and explicitness compared to non-translated native texts. Such features are interpreted as consequences of the higher than average cognitive effort associated with bilingual language activation, and possibly of an "overly strict conformance to perceived standard language norms" on the part of translators and non-native speakers (Kruger and van Rooy 2016: 27).

[1] Lanstyák and Heltai (2012) acknowledge that every form of communication is constrained in some way (physically, psychologically, socially, etc.). However, the authors use the term in a restricted sense to refer to communication where constraints "play a greater than average role" (p. 100).

The present study aims to contribute to this strand of research in two ways. First, it presents a corpus-based analysis of collocations, a lexical construct widely explored in research on translated and non-native language, but only touched upon in analyses comparing these constrained varieties to each other (see Section 2). Second, it compares, within a unified framework, written and spoken production by native and non-native speakers of English, as well as translations and interpretations into English of comparable texts. Building on Kruger's (2014) model of communication constraints, it thus aims to encompass all three dimensions which have been hypothesised to affect variation in constrained varieties, i.e. *modality* (written vs. spoken), *language activation* (monolingual vs. bilingual) and *text production* (mediated vs. unmediated).

The remainder of the article is structured as follows. Section 2 reviews previous work on collocations in constrained varieties, specifically on translation and non-native language. Section 3 presents the study setup, including the corpus used and the method adopted to extract and compare collocations across varieties, and Section 4 presents its results. Section 5 concludes by discussing results in the light of the literature on constrained communication, and offering a reflection on the inherent hurdles that corpus-based investigations face when dealing with language produced in contact settings.

2. Collocations in constrained language varieties

Conceptualising translated and non-native language through the lens of Kruger's (2014) model, one can see them as bilingual varieties distinguished by their text production status: the former is a mediated variety, i.e. one involving a mediation process between (source and target) texts, the latter is an unmediated one. Phraseological units, i.e. frequent or fixed lexical units involving more than one word, variously identified in the literature as collocations, formulae, idioms, etc. (see e.g. Gries 2008), have occupied centre stage in studies of these constrained varieties, though mostly separately.

In studies of translation and non-native language, phraseology is often the subject of comparisons between translated or non-native language on the one hand, and non-translated or native production on the other. The former varieties are hypothesised to display untypical patterning at the phraseological level compared to the latter, e.g. in terms of the word combinations used and/ or their frequency. In the case of translation, differences are interpreted against the background of purported translation norms or universals, such as cross-linguistic interference or normalisation (see Xiao and Hu (2015) for a recent overview). In non-native language, they are often taken as indicators of learner'/speakers' language competence (Gablasova, Brezina and McEnery 2017).

There are a multitude of approaches to the topic, so I will confine my attention to works dealing with the specific phraseological phenomenon that is focused on in the present study, i.e. collocations. Following Jones and Sinclair (1974/1996), these are defined as sequences of words that occur together more often than predicted by chance. Unlike qualitative approaches inspired by the Russian school of phraseology (Cowie 1998), which rely on criteria such as semantic non-compositionality and syntactic fixedness to define collocations (see, e.g., Nesselhauf 2005), this frequency-based view has the advantage of providing quantitative, replicable parameters for their identification in corpora. Reliance on objective criteria for collocation extraction, as Durrant and Schmitt (2009: 159) and Granger and Bestgen (2014: 240) observe, also allows for more straightforward comparisons of results across different studies.

No univocal trend emerges from studies focusing on collocations in translated language. Dayrell (2007), for example, finds that translations from Brazilian Portuguese into English display less variety in terms of collocations than comparable non-translated texts and Marco (2009) observes that collocations in English source texts are rendered with less well-established lexical units in Catalan, possibly as a result of the "translation difficulties often posed by phraseology" (p. 853). Results by Kenny (2001) and Bernardini (2011) point in the opposite direction: focusing on translations into English from German and Italian respectively, both authors observe a tendency on the part of translators to "opt for established collocations in the target language regardless of the presence of a corresponding collocation in the source text" (Bernardini 2011: 11), which they interpret as a signal of normalisation/standardisation (Toury 2004).

Concerning the use of collocations by non-native speakers/ learners, research on non-native language provides a clearer picture. According to Durrant and Schmitt (2009: 158), consistent evidence has emerged in the literature that learners, even at advanced levels, “use formulaic language […], but often not to the same extent as natives”, and that they “tend to overuse (in comparison to native norms) a small range of favourite phrases, especially if they are frequent/neutral items or are cognate to L1 forms”. The study by Durrant and Schmitt (2009) demonstrates the point with reference to different types of collocations, i.e. collocations characterised by high t-score values (t) vs. high Mutual Information values (MI). These are statistical association measures highlighting, respectively, high-frequency collocations (e.g. *year started*) vs. collocations that are less frequent but whose component words are highly associated to one another (e.g. *figures fluctuated*; see also Section 3.3). The authors find that learners of English tend to overuse high-t collocations and underuse high-MI ones compared to native speakers. Analogous findings are obtained by Granger and Bestgen (2014), who further suggest that use of high-MI collocations increases along with learner’s language proficiency; they also detect overuse of high-t collocations in the production of lower proficiency learners (on high-MI vs. high-frequency collocations in learner language see also O’Donnell, Römer and Ellis (2013)).

Two studies have focused on phraseology across different constrained varieties. Rabinovich et al. (2016) find support for the constrained language hypothesis with reference to English non-native and translated texts produced under the influence of different first languages; specifically, they observe fewer collocations in the constrained texts when compared to the unconstrained (native, untranslated) ones. It should be noted, however, that they make no attempt to distinguish collocations based on different association measures. Ferraresi and Miličević (2017) focus on translated and interpreted texts from English into Italian. These are seen as instances of bilingual, mediated varieties distinguished by their modality (written vs. oral), rather than their text production status (mediated vs. unmediated). In this case, the two constrained varieties are shown to differ substantially from each other: only high-t collocations are used to a similar extent in the two varieties, while high-MI collocations are significantly underused in interpretations, when compared to both translations and comparable original speeches.

Summing up, the research reviewed in this Section suggests that collocational patterns in constrained varieties, and more specifically in non-native and translated/interpreted language, generally differ from those in unconstrained varieties, though not always in similar ways. Even within single varieties, divergent tendencies have been observed depending on such factors as the language pairs and the nature of the collocations observed – to borrow the words of Toury (2004: 25), this is a "tangled knot", whose threads are not easily unravelled. By exploring whether and how different constraints affect the use of different types of collocations, the study presented in Section 3 aims to take a step in this direction, pointing to both shared and unique phraseological features across constrained varieties.

3. Comparing collocations in constrained English

3.1. Corpus

The corpus used in the study is an ad hoc subset of the *European Parliament Translation and Interpreting Corpus* (EPTIC; Bernardini, Ferraresi and Miličević 2016), a trilingual (English, French and Italian) corpus of European Parliament (EP) plenary speeches featuring four main components: (1) transcripts of speeches delivered at the EP plenary sessions, where speakers have the right to speak in their native language or in a language of their choice, including English; (2) verbatim written reports derived from these speeches; (3) transcripts of the simultaneous interpretations of the speeches; and (4) written translations of the verbatim reports, which result from an independently performed translation process, without any reference to the interpreters' outputs. All texts are annotated with part-of-speech and lemma information using the TreeTagger[2], and indexed for consultation through a NoSketch Engine platform (Rychlý 2007) hosted by the University of Bologna[3].

[2] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[3] https://corpora.dipintra.it/eptic/

Texts selected for the purposes of the present study include an equal number of original speeches delivered in English by native speakers (NS) and non-native speakers (NNS) (either read out or delivered impromptu), as well as the corresponding verbatim reports[4], and the interpretations and translations into English. The final dataset is shown in Table 1.

TABLE 1
Number of texts (and number of tokens) of the EPTIC corpus subset used in the study

	Unmediated			Mediated	
	Speeches		Reports	Interpreted	Translated
	Read-out	Impromptu			
NS	10 (3,082)	10 (2,111)	40 (11,534)	40 (11,891)	40 (12,168)
NNS	10 (4,904)	10 (1,871)			
TOTAL	40 (11,968)		40 (11,534)	40 (11,891)	40 (12,168)

While the corpus is very small and representative of a very peculiar communicative setting (the EP), comparability across all of its components is very high (see Section 5).

3.2. Study aims and setup

The present study compares a) *unconstrained varieties of English*, instantiated by unmediated oral speeches by NS and unmediated written reports; and b) *constrained varieties of English*, instantiated by oral speeches delivered by NNS, as well as interpretations and translations.

The study aims to answer two research questions:

– Do varieties characterised by different values along the three

[4] It was possible to distinguish between texts produced by NS and NNS only in the case of oral speeches, since the corresponding verbatim reports usually undergo an editing process before publication, hindering attempts at assessing the NS/NNS status of their ultimate producer (see Bernardini, Ferraresi and Miličević 2016: 68-69).

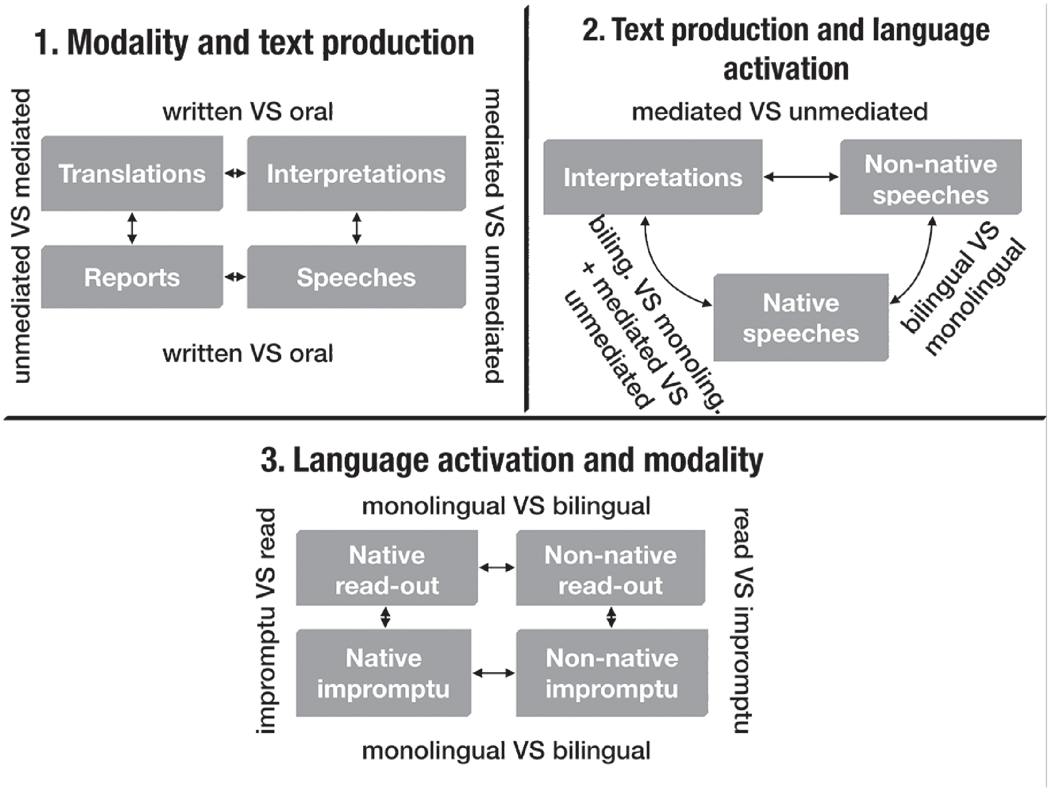
dimensions of constraint (modality, text production and language activation) display diverging patterns of collocation use?
– If so, do the bilingually constrained varieties display shared patterns that set them apart from the unconstrained ones?

These questions are addressed in three sets of comparisons focusing on the collocational parameters described in Section 3.3, with each set corresponding to one of the three possible pairwise permutations of Kruger's (2014) dimensions of constrainedness. It would have been ideal to test the combined effect of these dimensions within an all-in-one study design. Unfortunately, the corpus setup did not allow for this possibility, since not all possible combinations of values for all dimensions are represented in EPTIC – for instance, there are no written texts produced by NNS (see footnote 4).

The first set of comparisons is between translations, interpretations, unmediated written texts and unmediated oral speeches by NS, testing for the effect of modality (written/oral) and text production (mediated/unmediated) status. The second set of comparisons, which focuses on spoken data only, is between interpretations, speeches delivered by NNS and speeches delivered by NS, testing for the effect of text production (mediated/unmediated) and language activation (bilingual/monolingual) status. In the third set of comparisons, the effect of language activation (bilingual/monolingual) and modality (written/oral) status is examined in unmediated speeches by NS and NNS, taking into account their mode of delivery, i.e. impromptu versus read-out. While strictly speaking the modality of the latter category is spoken, it is hypothesised to share features with written language (Erman and Warren 2000), particularly at the lexical and phraseological levels[5]. Figure 1 illustrates the three sets of comparisons.

[5] It should be noted that, within the second analysis, read-out speeches are analysed jointly with the impromptu ones as instances of oral texts. This is done to increase the size of the sample. Since the conflation of read-out and impromptu speeches is performed both for the NS and NNS group, however, it should not introduce a bias in the comparisons focused upon in that analysis (mediated vs. unmediated and bilingual vs. monolingual production).

FIGURE 1
Comparisons carried out in the study



3.3. Method

Collocation candidates were extracted from the EPTIC sub-corpora based on the following part-of-speech patterns:

– adjective + noun: e.g. *political role*;
– noun + noun: e.g. *road safety*;
– verb + noun: e.g. *exploit (the) dimension*;
– noun + verb: e.g. *challenges arising*.

T-score and MI values were then computed relying on frequency data obtained from a large reference corpus of English (ukWaC; Baroni et al. 2009). The procedure, also implemented by Durrant and Schmitt (2009) and Granger and Bestgen (2014), has two advantages. First, it allows classification of collocation candidates based on their frequency in general English, rather than in the specific text types under scrutiny. Second, it overcomes the data sparseness problems often encountered when computing frequencies on the basis of very small corpora. As also mentioned in Section 2, the rationale for adopting these two association measures is that they highlight different types of collocation, i.e. frequent collocations often composed of frequent words in the case of t-score (e.g. *few groups*) vs. rarer collocations often composed of low-frequency words in the case of MI (e.g. *outspoken condemnation*). To increase the reliability of the measures, only pairs with frequency equal to or greater than 3 were retained for analysis (see Evert 2005).

Mean scores for all the extracted collocations were then calculated on a text-by-text basis for the two measures separately. In this way, two values were obtained for each text in each sub-corpus, providing an indication of the strength of its collocations as defined by t-score and MI values: the higher the scores, the stronger the collocations. In the final step, these mean scores were compared across sub-corpora within the three sets of comparisons outlined in Section 3.2, first by visual inspection of violin plots (see Section 4.1), and then by statistical testing. Testing was performed in R (R core team 2018) with one-way ANOVAs or Kruskal-Wallis tests, using the sub-corpus as a predictor variable; these were followed where appropriate by post-hoc pairwise comparisons using t-tests or Wilcoxon rank sum tests with Holm correction. The choice of parametric or non-parametric tests was based on preliminary checks on the normality of the distributions (Shapiro-Wilks tests), and the homogeneity of variance (Levene's tests). The results of the quantitative analyses are presented in Section 4.

4. Results

4.1. Mean MI values

Starting with the comparison of English varieties distinguished by modality and text production status, the violin plot in Figure 2 displays the distribution of mean MI values in the four sub-corpora under investigation. The width of the curved shapes is proportional to the number of texts whose collocations display similar mean values: the wider the shape, the higher the number of texts displaying a certain mean MI value. Boxplots are also drawn within these shapes to represent median values (visualised as thick white lines cutting the boxes in two) and the spread of the central data around them (within the first and third quartile, corresponding to the lower and upper lines of the boxes).

As Figure 2 shows, original speeches and written reports display higher MI values than the corresponding mediated varieties (interpretations and translations), and so do written texts when compared to their oral counterparts: written reports are "more collocational" than speeches, and translations are more collocational than interpretations.

Based on results of ANOVA, which reveals a significant difference among the four sub-corpora ($F(3,136)=8.79$, $p<0.001$, $\eta^2=0.162$), post-hoc t-tests were carried out to assess the significance of differences between pairs of sub-corpora (see Table 2). These reveal that only interpretations and speeches are significantly different from each other, hinting at an effect of text production status only in the oral mode.

FIGURE 2
Effect of modality and text production status

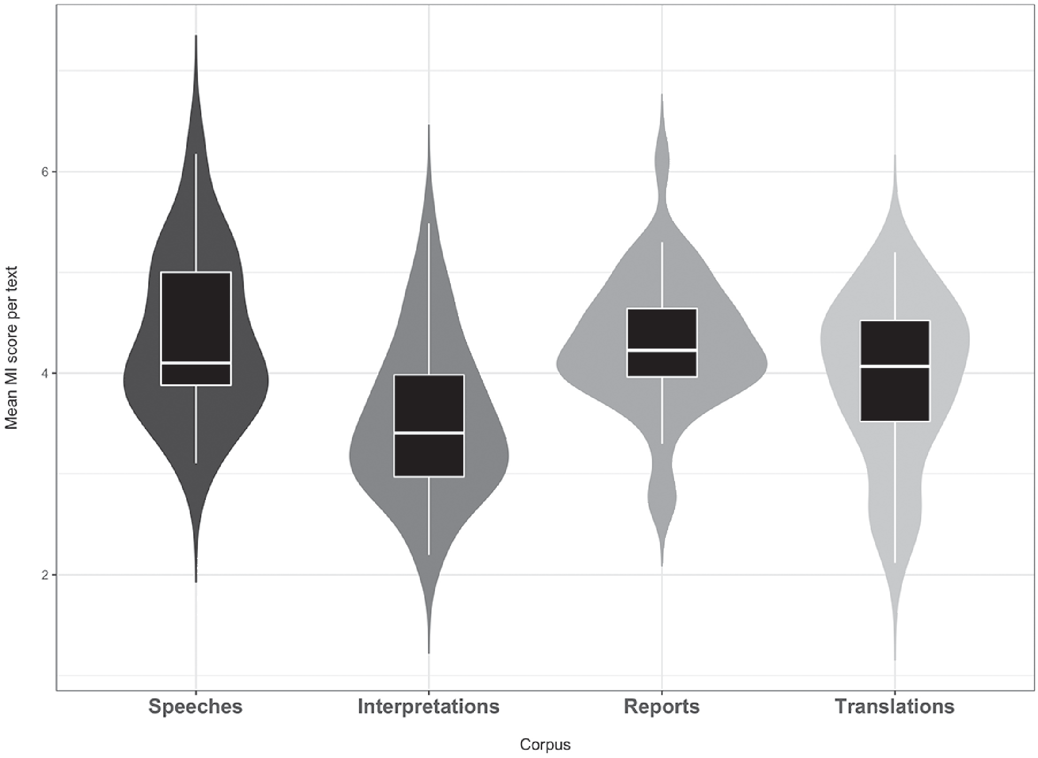
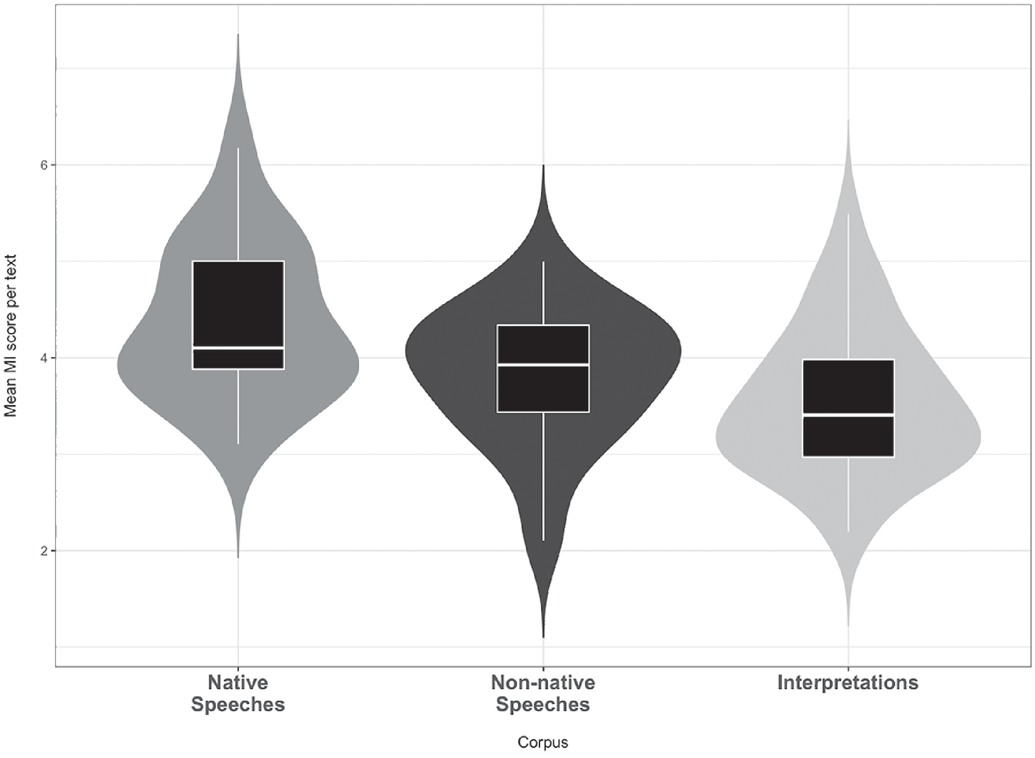


TABLE 2
Mean MI scores and SD values for sub-corpora involved in the modality and text production comparison (left part of the Table); p-values of pairwise t-tests with Holm correction (right part of the Table; significant result in bold)

	Mean MI	SD		Translations	Speeches
Speeches	4.3	0.8	Interpretations	0.07	**0.002**
Interpretations	3.5	0.7	Reports	0.07	0.72
Reports	4.4	0.6			
Translations	3.9	0.8			

Moving on, Figure 3 shows results for the second set of comparisons, focusing on text production and language activation status. Here, a cline in collocationality is observed going from the monolingual unmediated texts to the bilingual mediated ones: NS speeches display the highest mean values, interpretations display the lowest values, and the (bilingual unmediated) NNS speeches position themselves in-between.

FIGURE 3
Effect of text production and language activation status



The ANOVA points to a significant difference ($F(2,77)=7.89$, $p<0.001$, $\eta^2=0.172$), with post-hoc tests revealing a contrast between, on the one hand, NS speeches, and, on the other hand, NNS speeches and interpretations, which are not significantly different from each other (see Table 3). Hence, both text production and language activation are found to impact on collocation use, with the two constrained varieties being similar to each other and different from the unconstrained variety.

TABLE 3
Mean MI scores and SD values for sub-corpora involved in the text production and language activation comparison (left part of the Table); p-values of pairwise t-tests with Holm correction (right part of the Table; significant results in bold)

	Mean MI	SD		NNS speeches	Interpretations
NS speeches	4.3	0.8	NS speeches	**0.04**	**0.001**
NNS speeches	3.8	0.7	Interpret.	0.17	-
Interpretations	3.5	0.7			

Results of the third set of comparisons are presented in Figure 4. It will be remembered that while these analyses aim at detecting an effect of the language activation and modality constraints, they only involve spoken data, and specifically speeches delivered by NS and NNS in an impromptu vs. read-out modality (see Section 3.1). Focusing on modality differences, impromptu speeches display consistently lower collocationality values than the read-out ones. No clear tendency emerges when considering differences in terms of language activation: NS speeches tend to be more collocational than NNS ones, but this is mostly true of read-out speeches only.

As in the two previous analyses, the ANOVA returns significant results for comparisons across sub-corpora ($F(3,36)=3.7$, $p=0.02$, $\eta^2= 0.172$); yet no significant difference is observed in pairwise comparisons (see Tables 4a and 4b). No reliable effect of language activation and modality status can therefore be detected in this case. In interpreting this result, however, one should bear in mind that the sub-corpora involved in the analysis are minute, comprising ten texts each (see Section 3.2). In this light, one cannot rule out the possibility that lack of significance is due to the small sample size.

FIGURE 4
Effect of language activation and modality status

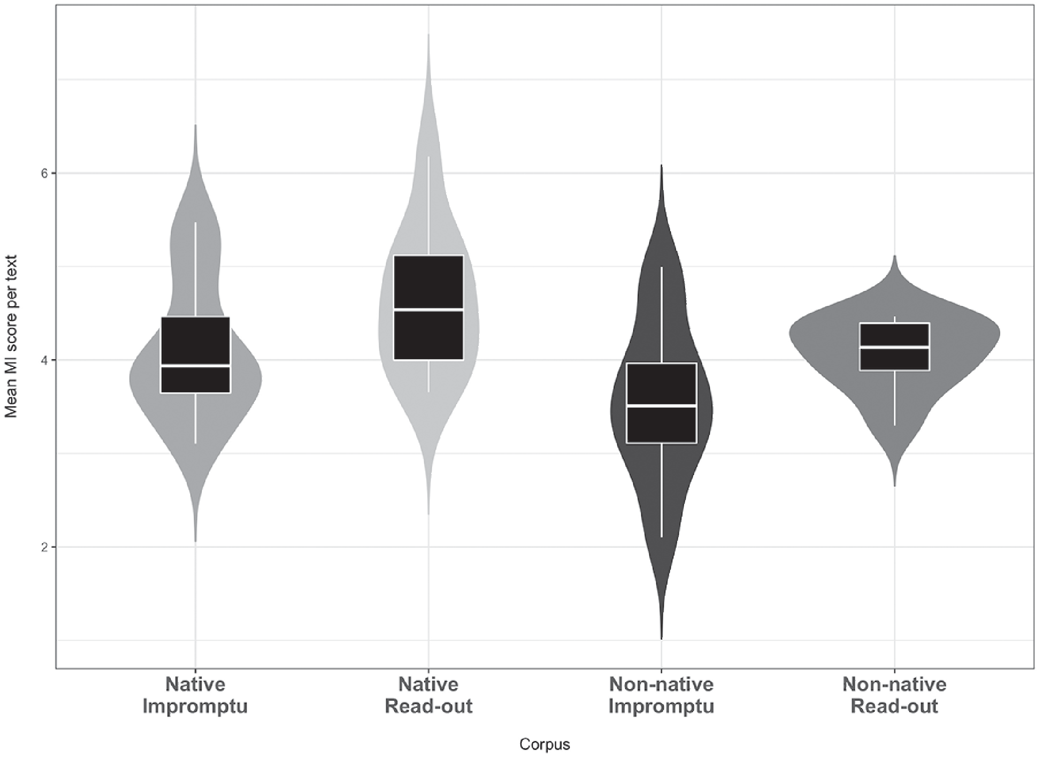


TABLE 4
Mean MI scores and SD values for sub-corpora involved in the language activation and modality comparison (left part of the Table); p-values of pairwise t-tests with Holm correction (right part of the Table)

	Mean MI	SD		NS impromptu	NNS read-out
NS read-out	4.1	0.8	NS read-out	0.27	0.22
NS impromptu	3.5	0.7	NNS impromptu	0.15	0.35
NNS read-out	4.3	0.6			
NNS impromptu	3.9	0.8			

4.2. Mean t-score values

Three Kruskal-Wallis tests mirroring the three ANOVAs presented in Section 4.1 were performed for t-score values, which displayed a non-normal distribution. None of the comparisons returned significant results:

- Comparison 1 (influence of modality and text production): H(3)=5.84, p=0.11;
- Comparison 2 (influence of text production and language activation): H(2)=2.56, p=0.27;
- Comparison 3 (influence of language activation and modality): H(3)=2.74, p=0.43.

For space reasons, neither full descriptive statistics nor plots for the comparisons are included. These, together with the full dataset used in the analysis, are available from the author upon request.

5. Discussion and conclusion

This contribution has presented a study focusing on collocations in a subset of the EPTIC corpus, encompassing English oral speeches by native and non-native speakers, their written-up versions, and their translations and interpretations, thus representing different values along the dimensions of communication constraints proposed by Kruger (2014). Against the backdrop of Lanstyák and Heltai's (2012) constrained language hypothesis, several sets of comparisons were carried out among these text sets, with the aim of identifying features typical of non-native and mediated (translated and interpreted) language.

Summing up the results presented in Section 4, the analysis of collocations defined by MI scores reveals that constraints related to language activation and text production affect use of collocational patterns, but only in the oral modality: both interpretations and spoken production by NNS display significantly lower collocational strength values than comparable spoken production by NS, but do not differ from each other. Unlike in the case of interpretations, translations are not significantly less collocational than their written unmediated counterpart, which advises against considering (comparatively) low MI values as a feature shared by all constrained varieties. The fact that no significant difference is observed between translations and interpretations, nor between original speeches and written-up versions, might then constitute further indication that, at least in these settings, the impact of modality-related constraints is more limited than that of other constraints.

In their investigation of translated and written non-native indigenised varieties of English, Kruger and van Rooy (2016: 26) found evidence of "a shared processing complexity effect, which [...] limits the use of less frequent, more marked grammatical options that are not easily accessible in situations of high cognitive demand". Results pertaining to MI-defined collocations obtained in the present study mirror these observations with reference to spoken, rather than written varieties, and in the area of lexis, rather than grammar. Psycholinguistic research has shown that low-frequency lexical sequences even if lexically cohesive, as in the case of high-MI collocations, are less easily retrieved from memory than high-frequency ones, and that this is especially true for on-line tasks and in NNS production (see e.g. Tremblay and Tucker 2011; Ellis, Simpson-Vlach and Maynard 2008). Like the underuse of marked grammatical options in Kruger and van Rooy's study, the comparatively low MI scores displayed by NNS oral production and interpreting are therefore likely manifestations of a "processing complexity effect", and one that is characteristic of *oral* constrained varieties.

Reflecting on the differences between translation and bilingualism, Lanstyák and Heltai (2012: 115) suggest that certain effects related to bilingual activation,

> [are] counterbalanced by the fact that translators have explicit knowledge of the rules and norms of the target language and may draw on this resource. They are conscious monitor users, and generally have more time to use the monitor than bilinguals.

If only speculatively, one might hypothesise that this "monitoring" effect can also explain why translated texts do not follow the pattern of lower collocationality displayed by the oral constrained varieties.

The same monitoring effect might also account for the overall lack of significant differences among varieties in terms of mean t-score values. As argued above, high-frequency collocations such as those with high t-score require fewer cognitive resources than low-frequency ones, and are therefore likely to be more easily available. Plevoets and Defrancq (2018: 2) observe that use of such high-frequency phraseology might in fact reduce cognitive load, especially "where processing demands exceed available capacity", as is the case in interpreting.

To conclude, we would like to go back one step to the non-significant comparisons among impromptu and read-out speeches by NS and NNS. Results of these comparisons were found to be consistent overall with those of other analyses pertaining to MI-defined collocations, with read-out speeches being more collocational than impromptu ones, and speeches by NS being more collocational than NNS ones (at least in the read-out mode): the less stringent the constraints on language production, the more collocational the texts. It was argued in Section 4.1, however, that no confident generalisation could be drawn from such results, mostly due to the very small size of the sample considered.

This illustrates a more general methodological issue which is likely to affect not only studies of bilingually constrained varieties, but any study investigating several dimensions of variation in a single design: observations are bound to be based on a limited number of texts, especially if these are comparable enough to allow multiple comparisons across oral and written, mediated and unmediated, native and non-native texts. Reaching compromises between corpus coverage and comparability of its components seems one of the main challenges that lie ahead.

References

Baroni, Marco, Bernardini, Silvia, Ferraresi, Adriano, Zanchetta, Eros, 2009, "The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-crawled Corpora", *Language Resources and Evaluation* 43 (3), pp. 209-26.

Bernardini, Silvia, 2011, "Monolingual Comparable Corpora and Parallel Corpora in the Search for Features of Translated Language", *Synaps* 26, pp. 2–13.

Bernardini, Silvia, Ferraresi, Adriano, Miličević, Maja, 2016, "From EPIC to EPTIC: Exploring Simplification in Interpreting and Translation from an Intermodal Perspective", *Target* 28 (1), pp. 61–86.

Cowie, Anthony P. (ed.), 1998, *Phraseology: Theory, Analysis, and Applications*, Oxford University Press, Oxford.

Dayrell, Carmen, 2007, “A Quantitative Approach to Compare Collocational Patterns in Translated and Non-translated Texts”, *International Journal of Corpus Linguistics* 12 (3), pp. 375–414.

Durrant, Philip and Schmitt, Norbert, 2009, “To What Extent do Native and Non-native Writers Make Use of Collocations?”, *International Review of Applied Linguistics* 47 (2), pp. 157–77.

Ellis, Nick C., Simpson-Vlach, Rita, Maynard, Carson, 2008, “Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL”, *TESOL Quarterly* 42 (3), pp. 375–96.

Erman, Britt and Warren, Beatrice, 2000, “The Idiom Principle and the Open Choice Principle”, *Speech* 20 (1), pp. 29–62.

Evert, Stefan, 2005, *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* University of Stuttgart doctoral thesis.

Ferraresi, Adriano and Miličević, Maja, 2017, “Phraseological Patterns in Interpreting and Translation: Similar or Different?”, In G. De Sutter, M.-A. Lefer and I. Delaere (eds.), *Empirical Translation Studies: New Methodological and Theoretical Traditions*, Walter de Gruyter, Berlin-Boston, pp. 157-82.

Gablasova, Dana, Brezina, Vaclav, McEnery, Tony, 2017, “Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence”, *Language Learning* 67 (1), pp. 155–79.

Granger, Sylviane and Bestgen, Yves, 2014, “The Use of Collocations by Intermediate Vs. Advanced Non-native Writers: A Bigram-based Study”, *International Review of Applied Linguistics* 52 (3), pp. 229–52.

Gries, Stefan Th., 2008, “Phraseology and Linguistic Theory. A Brief Survey” In S. Granger and F. Meunier (eds.), *Phraseology: An Interdisciplinary Perspective*, John Benjamins, Amsterdam-Philadelphia, 3–26.

Jarvis, Scott and Pavlenko, Aneta, 2008, *Crosslinguistic Influence in Language and Cognition*, Routledge, New York.

Jones, Susan and Sinclair, John M., 1974/1996, “English Lexical Collocations: A Study in Computational Linguistics”, in J. A. Foley (ed.), *Sinclair on lexis and lexicography*, UniPress, Singapore, pp. 21–54.

Kenny, Dorothy, 2001, *Lexis and Creativity in Translation. A Corpus-based Approach,* St. Jerome, Manchester.

Kolehmainen, Leena, Meriläinen, Lea, Riionheimo, Helka, 2014, “Interlingual Reduction: Evidence from Language Contacts, Translation and Second Language Acquisition”, in H. Paulasto, L. Meriläinen, H. Riionheimo and M. Kok (eds.), *Language Contacts at the Crossroads of Disciplines*, Cambridge Scholars Publishing, Newcastle upon Tyne, pp. 3–32.

Kranich, Svenja, 2014, “Translations as a Locus of Language Contact”, In J. House (ed.), *Translation: A Multidisciplinary Approach*. Palgrave Macmillan, London, pp. 96-115.

Kruger, Haidee, 2014, "Language Change, Photoshopped Language and Constrained Communication: Some New Ways of Thinking about Translation through Corpora", *EST Newsletter* 45, pp. 8–10.

Kruger, Haidee and van Rooy, Bertus, 2018, "Register Variation in Written Contact Varieties of English", *English World-Wide*, 39 (2), pp. 214–42.

Lanstyák, István and Heltai, Pál, 2012, "Universals in language contact and translation", *Across Languages and Cultures*, 13 (1), pp. 99–121.

Marco, Josep, 2009, "Normalisation and the Translation of Phraseology in the COVALT Corpus", *Meta* 54 (4), pp. 842–56.

Nesselhauf, Nadja, 2005, *Collocations in a Learner Corpus*, John Benjamins, Amsterdam.

O'Donnell, Matthew B., Römer, Ute, Ellis, Nick C., 2013, "The Development of Formulaic Sequences in First and Second Language Writing. Investigating Effects of Frequency, Association, and Native Norm", *International Journal of Corpus Linguistics* 18 (1), pp. 83–108.

Plevoets, Koen and Defrancq, Bart, 2018, "The Cognitive Load of Interpreters in the European Parliament: A Corpus-based Study of Predictors for the Disfluency *uh(m)*", *Interpreting* 20 (1), pp. 1–28.

R Core Team, 2018, *R: A Language and Environment for Statistical Computing,* R Foundation for Statistical Computing, Vienna.

Rabinovich, Ella, Nisioi, Sergiu, Ordan, Noam, Wintner, Shuly, 2016 "On the Similarities Between Native, Non-native and Translated Texts", in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Berlin, pp. 1870–881.

Rychlý, Pavel, 2007, "Manatee/Bonito – A Modular Corpus Manager", in *Proceedings of the 1st Workshop on Recent Advances in Slavonic Natural Language Processing*, Masaryk University, Brno, pp. 65-70.

Tirkkonen-Condit, Sonja, 2004, "Unique Items – Over- or Under-represented?", in A. Mauranen and P. Kujamäki (eds.), *Translation Universals – Do They Exist?*, John Benjamins, Amsterdam, pp. 177–184.

Toury, Gideon, 2004, "Probabilistic Explanations in Translation Studies", in A. Mauranen and P. Kujamäki (eds.), *Translation Universals – Do They Exist?*, John Benjamins, Amsterdam, pp. 15-32.

Tremblay, Antoine and Tucker, Benjamin V., 2011, "The Effects of N-gram Probabilistic Measures on the Recognition and Production of Four-word Sequences", *The Mental Lexicon* 6 (2), pp. 302–24.

Xiao, Richard and Hu, Xianyao, 2015, *Corpus-Based Studies of Translational Chinese in English-Chinese translation*, Springer, Heidelberg.