Looking for Traces of Mediation in Written Academic English

Giuseppe Palumbo

Abstract
Texts written in English by non-native speakers can be considered as instances of mediated language, where the mediation takes place between a writer's native language and English, seen, respectively, as the 'source' and 'target' poles. In investigating such texts, the methods of analysis can thus draw on assumptions and approaches used in both second language acquisition research and translation studies, starting from the idea that in mediated communication the target product can show traces of interference from features and traits associated with the source material. This study investigates a corpus of research articles written in English by authors with different language backgrounds in two academic disciplines (linguistics and agricultural economics). Focusing on the native language of writers, the investigation looks at part-of-speech (POS) distribution, understood as a marker and measure of linguistic distance (and a trace of mediation) between the texts. The analysis shows that differences in POS distribution can be interpreted primarily in terms of the authors' language backgrounds, but the discussion of findings extends to considering the specificities of linguistics and agricultural economics as discourse communities.

Keywords: ELF, nativeness, mediation, transfer, academic English.

1. Introduction

Having established a global dominance that sowed its seeds in the 1950s or perhaps even earlier (Gordin 2015), English is today the language of choice (or necessity) of most scientific and scholarly communities. Academic English has been studied extensively in both its spoken and written varieties, and several research approaches or paradigms can be said to have emerged (see the overview in Tribble 2017). The inherent multilingual background of much of today's academic communication has been explicitly considered in investigations emphasizing the *lingua franca* role of English (e.g. in Mauranen 2012). As regards written academic genres, the multilingual background of writers has mainly been taken into consideration in investigations of learner writing (e.g. Nesselhauf 2005) or in studies adopting a sociological perspective (e.g. Lillis and Curry 2010). Corpus-based investigations of academic language that consider the writers' language background have tended to focus on specific aspects such as formulaic language and lexical bundles (Salazar 2014; Esfandiari and Barbary 2017).

Texts written in English by non-native speakers can be considered as instances of mediated language, where the mediation takes place between a writer's native language and English, seen, respectively, as the 'source' and 'target' poles. In investigating such texts, the methods of analysis can thus draw on assumptions and approaches used in translation studies (as also suggested by Cook 2012), starting from the idea that in mediated communication the target product can show traces of interference from features and traits associated with the source material. A 'mediated communication' approach to investigating texts written in English by non-native speakers does not imply, per se, that prominence should be given to the (non-) nativeness factor. The native/non-native distinction can, however, be treated as "a useful heuristic" (Gnutzmann and Rabe 2014: 39) and contribute a fresh perspective on the study of how, in academic English, language and pragmatic norms are being re-negotiated in an international scenario. This article builds on a previous exploratory study (Palumbo 2017) in which a case was made for investigating written academic English using perspectives that take into account dimensions of language "contact" or "mediation". In that article, a crude quantitative corpus-based analysis was presented of morpho-syntactic and structural features of texts, based on simple descriptive statistics. That analysis is refined here based on the same corpus, this time introducing statistical tests aimed at providing firmer empirical support to quantitative findings, and in turn using these findings to identify areas worthy of further analysis.

2. A mediation perspective on written ELF

The study proposed in this article is an attempt to look at the *fabric*, or *make-up*, of texts written in English by writers with different language backgrounds. In this communication scenario, English is viewed as a contact language between people with different first languages, including native speakers of English. I concur with Jenkins (2014: 2) in considering this an English as a Lingua Franca (ELF) scenario; as pointed out in Mauranen (2018: 8), other definitions of ELF tend to exclude native speakers.

The general aim is to identify possible similarities and differences between the texts at the morpho-syntactic and structural level. The study will focus, in particular, on the distribution of parts of speech (POS) in the texts included in the corpus. POS distribution is understood as a possible marker of linguistic distance (and a trace of "mediation") between the texts. The study is designed so that, where differences and similarities between the corpus components are identified, they can be related to the authors' native languages, or families of languages. However, because the texts included in the corpus are from two different academic disciplines, attention will also be paid to the possible interaction of the lingua-cultural and disciplinary identities of the writers.

The analytical perspective adopted here in investigating written academic English draws explicitly on studies of transfer in second language acquisition research and studies of translated language that adopt a descriptive approach and try to characterize the linguistic make-up of translations against naturally-occurring non-translated texts. In transfer research, Jarvis (2000; 2010) has proposed a methodological framework that defines cross-linguistic influence as "a relationship between source language group membership and target-language behaviour" (Jarvis 2010: 170). The framework is aimed at identifying cross-linguistic effects, including those that "tend to go unnoticed because they are either subtle or unobservable without the use of proper analytical tools" (Jarvis 2010: 169). More specifically, the framework includes a scheme of comparison-based approaches that focus on the types of evidence and comparisons needed to identify transfer. In Jarvis (2010) four specific types of comparisons are identified using the two basic parameters of "group" (of speakers/writers) and "language" (L1 vs L2, or source vs target): intragroup homogeneity; intergroup heterogeneity; cross-language congruity; and intralingual contrasts. The first three of these comparisons are all likely to be relevant for the present study, which not only looks at possible evidence of transfer from different

L1s but also tries to apply a language-typological lens in discussing cross-linguistic effects.

In translation studies, the metaphor of “textual fit” has been proposed (originally by Chesterman 2004; see also Biel 2014) to describe how the language of translations differs from that of non-translations. One particular factor affecting the textual fit of translated language is interference, which is the subject of Toury’s (2012: 310) much-quoted “law of interference”: “in translation, phenomena pertaining to the make-up of the source text tend to force themselves on the translators and be transferred to the target text”. These phenomena can manifest themselves as “negative” or “positive” transfer, with transfer to be understood as the switching between source and target codes. Negative transfer derives from “deviations from codified practices in the target system” (Toury 2012: 311). Positive transfer results from “an increase in the frequency of features that the target system employs anyway” (Toury 2012: 311). A specific contribution to transfer research coming from translation studies may be found in the detailed classifications of typical or likely sources of transfer to be found in translation textbooks (such as Scarpa 2008). In terms of Jarvis’ (2010) framework mentioned above, such classifications might be construed as pools of empirically testable hypotheses about specific sources of transfer for given language pairs.

The assumption that the written output of non-native speakers of a given language can be discussed in terms of transfer from the writers’ respective native languages is not new in studies of written academic English. While most existing studies have focused on specific aspects of language, others have adopted a more general perspective. Rozycki and Johnson (2013), for instance, have discussed “non-canonical” grammar usage in a corpus of research articles, with “non-canonical” defined as usage that would be considered unacceptable by a standard grammar of the English language. In discussing their findings, they explicitly acknowledge the “carryover effect from the authors’ first language”, or “L1 interference”, while admitting that they “can only rarely untangle the [non-canonical] occurrences in terms of specific first-language effects”, mainly on account of the fact that several of the papers in their corpus are jointly authored by native and non-native speakers.

3. Corpus design and methods

The corpus under investigation is the same as the one used for the exploratory study in Palumbo (2017), which the present description follows very closely. The corpus was constructed on the basis of a single primary aim and according to a set of more detailed criteria following on from that aim. The primary aim was to reflect output in English as produced by two different groups of academic researchers: native speakers (NSs) of English on the one hand and non-native speakers (NNSs) of English on the other. The definition of "native speaker" for any given language is, as is well-known, not straightforward (Davies 2003) and may be next to impossible when specialist writing is concerned (Tribble 2017: 34-35). The criterion used to guarantee that the corpus would reflect the 'nativeness vs non-nativeness' distinction was an empirical one, based mainly on the authors' individual publishing histories. Authors based in an English-speaking country, affiliated to a research institution in that country and only publishing articles written in English were taken to be representative of "native" output. Authors based in a non-English speaking country, affiliated to a research institution there and having a history of publishing *both* in English and in another language (presumably, their "native" language) were taken to be representative of "non-native" writing in English.

Not all academic disciplines would easily lend themselves to a search for "non-native" authors as defined above. In several disciplines most, if not all, publishing for research purposes is in English, and therefore authors generally have no history of publishing texts belonging to the same genre (i.e. academic articles) in another language, even when they are based in a non-English speaking country. After some research, linguistics and agricultural economics were identified as two academic disciplines in which a substantial number of authors have a history of research-related publications in another language besides English.

A corpus with two component sub-corpora was then constructed: one each for linguistics (LING) and agricultural economics (AGRO). Each corpus component includes a total of 120 texts, all of them research articles, distributed as follows: 20 texts written by native speakers of English; plus another 100 texts written by non-native speakers of English, with 20 texts from each of 5 different native-language backgrounds, namely Croatian, German, Italian, Polish and Spanish. Each corpus component thus comprises six different native-language (NL) sets, English being one of these native languages (see Table 1). For the non-native English-speaking authors, the native languages were chosen partly for opportunistic reasons (i.e. because for those languages it was easier to find texts written by authors meeting the requirements described above), and partly with the aim of representing a variety of language families, so that results from the analysis could also be viewed in language-typological terms. Three language families are represented in the corpus: Germanic (English, German), Romance (Italian, Spanish) and Slavic (Croatian, Polish).

TABLE 1
Corpus composition and data

LING	**AGRO**
120 research articles in linguistics, representing 6 different native languages (20 texts per native language).	120 research articles in agricultural economics, representing 6 different native languages (20 texts per native language).
Total size: 581,100 words	Total size: 571,719 words
No. of words per language:	No. of words per language:
English – 104,505	English – 119,960
German – 103,605	German – 115,242
Croatian – 92,340	Croatian – 64,470
Polish – 95,930	Polish – 82,234
Italian – 92,280	Italian – 104,000
Spanish – 92,440	Spanish – 85,813

In order to meet the (non-)nativeness criteria illustrated above, compromises had to be made in terms of text selection. Most of the articles included in the corpus were taken from journals, with a preference for those adopting an open-access policy. In the AGRO corpus, however, a large number (i.e. around 50) of the selected articles across the various NL sets appear in conference proceedings or are available online as “papers prepared for publication” or “working papers”. In LING, it was possible to select single-authored articles in the vast majority of cases, whereas a considerable proportion of the articles in AGRO are authored by two (and sometimes three) researchers – all of whom were checked for their shared NL background. All texts were considered in their final, published form. As regards the time span for publication, both corpus components include articles published between 2005 and 2017.

Before inclusion in the corpus all the texts were cleaned by removing (most) para-textual material. The corpus was compiled and POS-tagged using the Sketch Engine (Kilgarriff et al. 2014) and the Penn Treebank tagset (older version[1]).

The analysis was carried out in two steps. The first step was quantitative. Figures on the frequency distribution of POSs across the NL sets in each sub-corpus were obtained and they were then submitted to statistical analysis, so as to identify the POS categories for which significant differences in frequency distribution emerged across the NL sets. More specifically, for each sub-corpus non-parametric tests were used to establish whether the differences observed in the distribution of POS categories across the NL sets could be deemed significant. The tests were conducted based on paired NL sets, with the pairings decided on the basis of language family affiliations: Croatian-Polish (CR-PL), Italian-Spanish (IT-SP) and English-German (EN-DE).

The second analytical step was of an essentially qualitative nature. The focus was this time on the POS categories for which statistically significant differences emerged across the paired NL sets. In particular, an attempt was made to interpret such differences as effects of the language backgrounds of the writers or, more specifically, as consequences of interference from the authors' native languages.

4. Results and discussion

In both subcorpora, the frequency of each POS category was first calculated for each individual NL set. Although a total of 52 POS

[1] The complete POS tagset is available on the Sketch Engine website at https://www.sketchengine.eu/penn-treebank-tagset/ – in the present study special characters (i.e. "#" and "$") and punctuation marks were excluded from the frequency counts and the subsequent statistical analysis, with the exception of the tag SENT, which identifies sentence-break punctuation marks (i.e. ".", "!" and "?").

categories was considered, I will not give a complete breakdown of the frequency counts for space reasons. Tables 2 and 3 give a simplified breakdown of more general, cumulative POS 'classes' in each sub-corpus, namely nouns, verbs (including modals), determiners, adjectives, adverbs, coordinating conjunctions, and prepositions and subordinating conjunctions (for these last two parts of speech the automatic POS annotation in the Sketch Engine has a single category). Due to the way counts are carried out by the software, the figures resulting from POS counts do not add up to the total number of running words.

TABLE 2
Relative proportions (in %) of cumulative POS classes in AGRO, per NL set

	CR	DE	EN	IT	PL	SP
	N=70672	N=121400	N=129329	N=112033	N=89317	N=94082
Nouns	28.44	27.65	25.85	25.15	26.68	26.42
Verbs	11.29	13.08	14.12	13.24	12.30	12.68
Adverbs	3.46	4.13	4.27	3.61	3.49	3.05
Determiners	8.06	10.42	9.72	11.49	10.94	10.55
Coordinating conj.	4.39	3.47	3.95	3.85	3.40	3.65
Prepositions/sub. conj.	15.10	14.65	13.57	14.08	15.49	13.71

A few remarks can already be made on the basis of the very general picture provided by the two tables. Overall, AGRO seems to be more 'nouny' than LING: the relative proportions of nouns in the former are all consistently higher than those in the latter. LING, on the other hand, displays a consistently higher proportion of verbs for all NL sets, and no single NL set clearly stands out from the others. This is not true of AGRO, where EN is the NL set with the highest relative proportion of verbs – a finding which resonates with descriptions of English as a language that favours verb-based constructions, at least in comparison with Romance languages (as is often discussed in studies of specialized translation, such as Scarpa 2008 and Musacchio 2017). Even in such a general picture, a distinction seems to emerge between the two sub-corpora, with one of them (i.e. LING) appearing to be more uniform across the different NL sets in the way parts of speech are distributed. Yet even in LING there are categories for which the NL sets display marked variation, as happens with coordinating conjunctions, for which the relative proportion in EN is markedly higher than in the other NL sets.

TABLE 3
Relative proportions (in %) of cumulative POS classes in LING, per NL set

	CR	DE	EN	IT	PL	SP
	N=99703	N=110294	N=111341	N=98616	N=103791	N=99375
Nouns	22.98	23.79	24.00	23.37	24.01	23.66
Verbs	14.23	14.84	14.48	14.51	14.96	13.37
Adverbs	4.64	5.08	4.91	4.81	4.46	4.42
Determiners	10.88	10.66	9.58	10.59	10.16	11.43
Coordinating conj.	3.19	3.23	4.06	3.59	3.23	3.47
Prepositions/sub. conj.	13.99	13.85	13.67	14.48	13.39	14.17

To enhance the granularity of the analysis and, more importantly, to identify cases in which differences across the NL sets could be observed to be statistically significant, the frequency counts for all the individual POS categories were submitted to non-parametric tests using the Stata 11 software (StataCorp 2009). A non-parametric Kruskall-Wallis test ($p<0.05$) was first used to compare three pairings of NL sets, with the pairings decided on the basis of language family affiliations: Croatian-Polish (CR-PL), Italian-Spanish (IT-SP) and English-German (EN-DE). Subsequently, comparisons between groups (IT-SP versus EN-DE; CR-PL versus EN-DE; CR-PL versus IT-SP) were evaluated through a non-parametric Dunn's test ($p<0.05$) with Bonferroni correction for multiple comparisons, so as to lower the risk of obtaining significant results by chance. It should be noted that the study employs measures of statistical significance not so much as a means to identify key items, but rather as a way to facilitate comparisons across paired NL sets. As discussed in Gabrielatos (2018), statistical significance scores may not accurately reflect the size of a frequency difference, and keyness may have to be established through an effect-size metric.

For each sub-corpus, POS categories were identified that yielded significant differences in both tests. Table 4 lists such categories and also provides the following information: the median number of occurrences of a POS in each pair of NL sets; the actual p value obtained from the analysis of paired NL sets; and an indication of the specific comparison(s) that emerged as significant from the Bonferroni-corrected analysis. Although no effect-size metric is employed, the indication of median occurrences is intended to provide an approximate measure of the size of frequency differences across the paired NL sets.

So, for example, for the category of coordinating conjunctions (CC) in AGRO, p equals 0.0005 and the median occurrence of CC tokens in each pair of NL sets is as follows: 145 for CR-PL, 217 for DE-EN, and 191 for IT-PL. In LING, p equals 0.0131 and the median number of occurrences is 161 for CR-PL, 186 for DE-EN, and 168.5 for IT-PL. For both AGRO and LING, one specific pair-based comparison (that between CR-PL and EN-DE) also emerged as significant in this category. The CC category is the one containing items such as *and*, *but*, *both*, *either*, and *yet*. What the figures are telling us is that, in both sub-corpora, such items are used by English and German authors the most and by Croatian and Polish authors the least, with Italian and Spanish authors coming in between.

On the basis of these results, the hypothesis is put forward that the native language of the authors is a factor influencing the choices made by writers in constructing, morpho-syntactically, their texts in English – a factor that ‘shines through’ the texts even in spite of the editorial process the texts (may) have undergone. To go back to the mediation perspective outlined in section 2, the EN native-language sets in the corpus are here being taken as those representing “canonical” usage (Roycki and Johnson 2013) or, in translation-studies terms, as the ‘non-translated’ or ‘target’ pole. The figures in Table 4 can thus be taken to provide us with a measure of how distant the authors with a non-native language background are from the native-language authors in the morpho-syntactic construction of their texts. The measure can be based on both the p value and the median of occurrences: the lower the p value and the wider the spread between the number of occurrences, the more significant the distance between the pairs of NL sets.

As shown in Table 4, the number of POS categories testing for significant differences is much higher in AGRO than in LING, which confirms the trend observed when discussing Tables 2 and 3. Of the 29 POS categories listed in the table, only ten tested significant in both AGRO and LING (also note that no category tested significant only in LING): coordinating conjunctions, existential *there*, *that* as subordinator, singular or mass nouns, adverbs, sentence-break punctuation, singular present non-3rd person *be* (i.e. the word *are*), and verbs in the present tense (both 3rd and non-3rd person; that is, the categories labelled as VVP and VVZ). This suggests that, while the authors' language background is a significant factor in contributing to the linguistic make-up of texts, something else must be at work that determines differences between the NL sets. Hypotheses in this regard are bound to be highly speculative in the present context, but factors connected with the design of the corpus and the academic disciplines represented in it may shed some light on this difference in results between the two sub-corpora.

Recall, first of all, that in order to meet the stringent (non-) nativeness criteria for inclusion in the corpus, the articles for AGRO had to be selected from a variety of sources: not only journals and conference proceedings, but also websites where the articles had been posted as "working papers" or "papers prepared for publication". In LING, all the articles except one were taken from journals – the one exception being a text appearing in an edited collection. Because journals are the exclusive source for the articles in LING, this may have ensured heightened attention to language. It is conceivable, in other words, that most of the texts went through an editorial process which led to changes in the English originally used by the authors (thus resulting in more convergence on "canonical" usage). In AGRO, on the other hand, a significant number of the selected texts may have been made publicly available without any editorial interventions. The possibility that some of these texts are the result of translations from original foreign language drafts should not be discarded either.

TABLE 4
POS categories testing for significant differences across paired NL sets (for symbols, see legend below)

	AGRO					LING				
Part of speech	Median no. of occurrences			p-value	Multiple comparisons	Median no. of occurrences			p-value	Multiple comparisons
(with indication of Sketch Engine tag)	CR-PL	EN-DE	IT-SP			CR-PL	EN-DE	IT-SP		
coordinating conjunctions (CC)	145	217	191	0.0005	(a)	161	186	168.5	0.0131	(a)
existential *there* (EX)	7	12	8	0.0215	(a)	8	13	7	0.0080	(c)
determiners (DET)	359	632	516	0.0001	(a) (b)	542	552	545.5	ns	-
preposition, subordinating conjunction (IN)	607	786	651.5	0.0006	(a)	727.5	748.5	733	ns	-
that as subordinator (IN/that)	14	30.5	17	0.0004	(a)	38	41.5	31	0.0006	(b) (c)
adjectives (JJ)	359	572	443.5	0.0001	(a) (c)	509.5	508.5	484.5	ns	-
adjectives, comparative form (JJR)	16	24.5	16	0.0104	(a)	16.5	18	13	ns	-
modals (MD)	26.5	58	37	0.0001	(a) (b) (c)	46.5	52	43	ns	-
nouns, singular or mass (NN)	703.5	1099	815	0.0001	(a) (c)	851	903	799.5	0.0015	(a) (c)
nouns, plural (NNS)	331	522	410	0.0001	(a) (b) (c)	388.5	403	359.5	ns	-
personal pronouns (PP)	24.5	49	37	0.0033	(a)	78.5	86	76	ns	-

(*continued on next page*)

TABLE 4 (*continued from previous page*)
POS categories testing for significant differences across paired NL sets (for symbols, see legend below)

possessive pronouns (PPZ)	16	25	24	0.0237	(a)	36.5	39	35.5	ns	-
adverbs (RB)	98.5	235	124	0.0001	(a) (c)	203	239	205	0.0223	(a) (c)
adverbs, comparative (RBR)	8	17	9	0.0113	(a)	14	16	14	ns	-
infinitive *to* (TO)	27	77	57	0.0001	(a) (b) (c)	64	64	61	ns	-
sentence-break punctuation (SENT)	130.5	236.5	177	0.0001	(a) (b) (c)	215	218	171	0.0001	(b) (c)
verb *be*, base form (VB)	17.5	41.5	23	0.0001	(a) (c)	32	36	36	ns	-
verb *be*, past participle (VBN)	6.5	11	10	0.0111	(a) (b)	10	10	10	ns	-
verb *be*, sing. pres. non-3rd (VBP)	23	46	38	0.0001	(a) (b) (c)	31	42	30	0.0176	(a) (c)
verb *be*, 3rd person sing. present (VBZ)	44	69.5	53	0.0008	(a) (c)	65.5	78	71.5	ns	-
verb *have*, non 3rd person, present (VHP)	8	16	11	0.0082	(a)	9	11	11	ns	-
verb *have*, 3rd person sing., present (VHZ)	9	15	13.5	0.0039	(a) (b)	10	13	10.5	0.0252	(a)
verb, base form (VV)	43	122.5	79	0.0001	(a) (b) (c)	99.5	99.5	89.5	ns	-

(*continued on next page*)

TABLE 4 (*continued from previous page*)
POS categories testing for significant differences across paired NL sets (for symbols, see legend below)

verb, gerund/present participle (VVG)	72	94	81	0.0139	(a)	90.5	85	82	ns	-
verb, past participle (VVN)	97	160.5	127	0.0001	(a) (b) (c)	157	149.5	154	ns	-
verb, present, non-3d (VVP)	15	46	34	0.0001	(a) (b) (c)	44	49.5	41	0.0225	(c)
verb, 3rd person sing. present (VVZ)	27	58	42.5	0.0001	(a) (b)	57	69	56	0.0272	(c)
WH-determiner (WDT)	18.5	35	29	0.0001	(a) (b)	38.5	44	39	ns	-
WH-adverb (WRB)	8	18	10.5	0.0006	(a) (c)	15.5	16.5	18.5	ns	-

Key:
ns = not significant;
(a) = comparison (CR-PL) vs (EN-DE);
(b) = comparison (CR-PL) vs (IT-SP);
(c) = comparison (IT-SP) vs (EN-DE).

The nature of the two represented disciplines may also have played a role in determining the observed differences. Researchers in linguistics, especially those for whom English is a subject of research, can be assumed to have a better command of English, leading them to draft texts that conform more closely to canonical usages. For the same reason, both the authors and the editorial gatekeepers could be assumed to exhibit a lower level of tolerance for deviant usage. Attitudes to English as the medium of research dissemination are perhaps different in a discipline such as agricultural economics. Based on their own findings from a corpus of computer engineering articles, Rozycki and Johnson (2013) conclude that "engineering gatekeepers exhibit a willingness to accommodate [non canonical] usage, and readers appear willing to negotiate the meaning of texts with [non canonical] usage". Similar attitudes may have emerged in agricultural economics as well. Tolerance for deviant English language usage, in other words, may be markedly different across disciplines. In translation studies terms, an echo of this is Toury's (2012; 314) idea that tolerance (and therefore the realization) of interference varies according to the socio-cultural conditions in which translations take place.

4.1. A small illustrative case study

Analysing all the factors that may have resulted in significant quantitative differences for each of the POS categories in Table 4 is beyond the scope of the present study. But I will now focus on two POS categories, which will serve as a small illustrative case study of identification of aspects and methods of investigation that could serve as a template for the study of the other categories.

The two selected categories are 'verb *be*, sing. pres. non-3rd' (VBP) and 'verb *be*, 3rd person sing. present' (VBZ) – in other words, the two present-tense forms of the verb 'to be': *is* and *are*. The first category tested significant in both LING and AGRO, while the second tested significant only in AGRO. I am considering them together because, given their closeness, they may be expected to give rise to similar colligational/collocational patterns. The statistical analysis tells us that the two forms are the most frequent in the EN and DE sets in both disciplinary components. One intuitive expectation could be that the higher frequency of occurrence is linked to a more frequent use of passive constructions by the English and German authors. Table 5 gives the cumulative number of occurrences and the relative frequency (per thousand words) of *is* and *are* in AGRO, for each individual NL set. The table also specifies the number of times *is/|are* is followed by a past participle and the number of times *is|/are* is followed by a different part of speech.

TABLE 5
Absolute and relative frequency (per 100k words) of *is|are* in AGRO, per NL set

	CR	DE	EN	IT	PL	SP
is|are total	1429 (1822.9)	2457 (1815.6)	2632 (1854.3)	2115 (1687.5)	1427 (1453.1)	1693 (1639.7)
is|are + past participle	356 (454.1)	755 (557.9)	531 (374.1)	518 (413,3)	390 (397.1)	480 (464.9)
is|are + other POS	1073 (1368.9)	1702 (1257.7)	2101 (1480.2)	1597 (1274,2)	1037 (1056)	1213 (1174.8)

Contrary to expectations, although English authors cumulatively use the two items *is* and *are* the most, they are *not* the heaviest users of passive constructions — the heaviest users of passives are the authors in the German NL set. A closer look at the other POS appearing next to *is/are* helps to clarify what other patterns make them particularly frequent in EN. Among the top five POS collocates to the right of *is/are* in EN, *-ing* forms feature prominently, with 115 occurrences: in particular, *is/are going* and *is/are becoming* are both much more common in EN than in the other NL sets. To sum up, the high frequency of the present-tense forms of the verb 'to be' in this NL set is due to the frequency of passives as much as to the frequency of progressive forms, which is not true of the other NL sets. Similar analyses focusing on colligational and collocational patterns could be carried out for the other POSs so as to uncover the structures that lead to the significant differences in distribution emerging from the statistical analysis.

5. Conclusions and future research

The analysis reported in this contribution shows that adopting a 'mediation' approach to investigating texts written in English by authors with diverse language backgrounds can contribute to uncovering some subtle differences in the linguistic make-up of the texts. At the same time, the discussion has shown that, even when texts are observed primarily as "systems of forms" (Hyland 2016), some findings still need to be interpreted through a discourse perspective.

The investigation has also shown that future work could profitably focus on colligational and collocational patterns. Some studies of translated language have already elaborated on this idea, taking into consideration POS n-grams and their positional differences within texts (as in Borin and Prütz 2001) or representing text documents as "feature vectors" (Baroni and Bernardini 2006) which encode units of varying sizes (i.e. not only unigrams) and types (i.e. not only POSs but also lemmas and word forms). Similar approaches could be attempted on the corpus investigated here, serving as further empirical testing of the profitability of a 'mediation' approach to studying ELF texts.

Acknowledgements

I am deeply indebted to Antonino Bella, a researcher with Italy's *Istituto Superiore di Sanità*, for his help on the statistical analyses presented in this article. Any inaccuracies or inconsistencies in the interpretation of results are entirely my own. I am also extremely grateful to the two anonymous reviewers of an early draft of the paper for their many insightful comments and suggestions.

References

Baroni, Marco and Bernardini, Silvia, 2006, "A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text", *Literary and Linguistic Computing* 21(3), pp. 259-274.

Borin, Lars and Prütz, Klas, 2001, "Through a Glass Darkly: Part-of-speech Distribution in Original and Translated Text", in W. Daelemans, K. Sima'an, J. Veenstra and J. Zavrel (eds) *Computational Linguistics in the Netherlands 2000*, Rodopi, Amsterdam, pp. 30-44.

Biel, Łucja, 2014, *Lost in the Eurofog: The Textual Fit of Translated Law*, Peter Lang, Frankfurt am Main.

Chesterman, Andrew, 2004, "Hypotheses About Translation Universals", in G. Hansen, K. Malmkjær and D. Gile (eds), *Claims, Changes and Challenges in Translation Studies*, John Benjamins, Amsterdam.

Cook, Guy, 2012, "ELF and Translation and Interpreting: Common Ground, Common Interest, Common Cause, in *Journal of English as a Lingua Franca* 1(2), pp. 241-62.

Davies, Alan, 2003, *The Native Speaker. Myth and Reality*, 2nd ed., Multilingual Matters, Clevedon.

Esfandiari, Rajab and Barbary, Fatima, 2017, "A Contrastive Corpus-driven Study of Lexical Bundles between English Writers and Persian Writers in Psychology Research Articles", *Journal of English for Academic Purposes* 29, pp. 21-42.

Gabrielatos, Costas, 2018, "Keyness Analysis: Nature, Metrics and Techniques", in C. Taylor and A. Marchi (eds) *Corpus Approaches to Discourse: A Critical Review*, Routledge, London-New York, pp. 228-58.

Gnutzmann, Claus and Rabe, Frank, 2014, "'Theoretical Subtleties' or 'Text Modules'? German Researchers' Language Demands and Attitudes across Disciplinary Cultures", *Journal of English for Academic Purposes* 13, pp. 31-40.

Gordin, Michael D., 2015, *Scientific Babel. The Language of Science from the Fall of Latin to the Rise of English*, The University of Chicago Press, Chicago.

Jarvis, Scott, 2000, "Methodological Rigor in the Study of Transfer: Identifying L1 Influence in the Interlanguage Lexicon", *Language Learning* 34(2), pp. 1-17.

Jarvis, Scott, 2010, "Comparison-based and Detection-based Approaches to Transfer Research", *EUROSLA Yearbook* 10, 169-92.

Jenkins, Jennifer, 2014, *English as a Lingua Franca in the International University: The Politics of Academic English Language Policy*, Routledge, Abingdon.

Hyland, Ken, 2016, "Methods and Methodologies in Second Language Writing Research", *System* 59, pp. 116-25.

Kilgarriff, Adam, Baisa, Vit., Bušta, Jan, Jakubíček, Miloš, Kovář, Vojtěch, Michelfeit, Jan, Rychlý Pavel, Suchomel, Vit, 2014, "The Sketch Engine: Ten Years on", *Lexicography* 1(1), pp. 7-36.

Lillis, Theresa M. and Curry, Mary Jane, 2010, *Academic Writing in a Global Context. The Politics and Practices of Publishing in English*, Routledge, London-New York.

Mauranen, Anna, 2012, *Exploring ELF: Academic English Shaped by Non-native Speakers*, CUP, Cambridge.

MAURANEN, ANNA, 2018, "Conceptualising ELF", in J. Jenkins, W. Baker and M. Dewey (eds) *The Routledge Handbook of English as a Lingua Franca*, Routledge, London-New York, pp. 7-24.

MUSACCHIO, MARIA TERESA, 2017, *Translating Popular Science*, CLEUP, Padova.

NESSELHAUF, NADJA, 2005, *Collocations in a Learner Corpus*, John Benjamins, Amsterdam.

PALUMBO, GIUSEPPE, 2017, "Notes on Investigating the Native vs Non-native Distinction in Written Academic English", in G. Palumbo (ed.), *Testi, corpora, confronti interlinguistici: approcci qualitativi e quantitativi*, EUT Edizioni Università di Trieste, Trieste, pp. 111-25.

ROZYCKI, WILLIAM and JOHNSON, NEIL H., 2013, "Non-canonical Grammar in Best Paper Award Winners in Engineering", *English for Specific Purposes* 32(3), pp. 57-169

SALAZAR, DANICA, 2014, *Lexical Bundles in Native and Non-Native Scientific Writing: Applying A Corpus-Based Study to Language Teaching*, John Benjamins, Amsterdam.

SCARPA, FEDERICA, 2008, *La traduzione specializzata. Un approccio didattico professionale*, 2nd ed., Hoepli, Milano.

STATACORP, 2009, *Stata Statistical Software*: Release 11, College Station, TX, StataCorp LP.

TOURY, GIDEON, 2012, *Descriptive Translation Studies – and Beyond*, revised edition, John Benjamins, Amsterdam.

TRIBBLE, CHRISTOPHER, 2017, "ELFA vs. Genre: A New Paradigm War in EAP Writing Instruction?", *Journal of English for Academic Purposes* 25, pp. 30-44.