Predicting Voice Alternation Across Academic Englishes

1. Introduction

One of the aims of the *International Corpus of English* (ICE)[1] is to enable comparative studies of variation across a broad range of World Englishes (WEs). Initially, research was somewhat limited by the fact that most ICE corpora lacked grammatical annotation. This is particularly problematic for the study of relatively abstract grammatical patterns, such as voice. Even in a part-of-speech-tagged corpus, it remains difficult to efficiently retrieve active transitive counterparts of passives, making principled study of the active-passive alternation across WEs practically impossible. Now that a large number of ICE components have been parsed, automatic retrieval of actives and passives is possible.

Voice alternation is of interest because previous research (e.g. Leech et al. 2009) identified regional differences in a shift from passive towards active constructions, which is more pronounced in American (AmE) than in British English (BrE) academic writing. The question was whether the trend towards greater use of the active voice could also be found in other first- and second-language varieties of English (see Hundt et al. 2016). In this paper, we add to previous research by looking into the factors predicting the choice between active and passive. In particular, we aim to discover whether or not WEs share the core grammar of voice alternation. Substrate influence in academic writing – a highly edited register – will not be the same as in spontaneous spoken language where we find evidence of obvious structural transfer or borrowing, e.g. in the form of the *kena* passive in Singapore English (Bao & Wee 1999). All the same, substrate may play a role at a more subtle level, e.g. in a variety's readiness to combine passive voice with marking for other verbal categories such as aspect or giving preference to inanimate subjects (see section 2.4).

In part two of the paper, we provide background on previous research, the

[1] ICE samples acrolectal/standard English spoken as a first or institutionalised second language. For more background information, see the project website at [URL will be provided once the new website has been announced].

rationale for the choice of varieties we investigate, and on passives in the substrate languages. Section three gives information on data retrieval and coding of factors. Section four reports the results of our statistical analysis, which are discussed in the context of modelling probabilistic variation across WEs (section five).

2. Background

2.1 Previous research: *be*-passives across time and space

The overall frequency of *be*-passives varies across time and space: the long-term trend towards declining use of passives is more advanced in AmE than in BrE (Biber & Finegan 1989; Seoane 2006; Leech et al. 2009). On the basis of manual analysis, earlier studies show that this decline happens at the expense of active transitives (Seoane & Loureiro-Porto 2005) in academic writing, and that the trend is more pronounced in natural sciences than in the humanities (Hundt & Mair 1999). Research on the passive in second-language varieties of English is scarce (but see, e.g., Biewer 2009).

Hundt et al.'s (2016) analysis of the academic section of the parsed Brown-family of corpora and 15 ICE components corroborates previous results. Somewhat surprisingly, however, the broader selection of WEs did not reveal a divide into first- (ENL) and second-language (ESL) varieties. Instead, AmE was the only variety that significantly differed from other WEs in preferring actives in academic writing. Regression analysis showed that variation across sub-disciplines (humanities and social sciences vs. natural sciences and technology) was even more pronounced than regional variation, with the 'soft sciences' preferring a more active style (see also Biber & Finegan 1989). In other words, despite substantial differences in the substrate languages, stylistic preferences in the various sub-disciplines turned out to be the decisive factor for the overall frequencies of *be*-passives and active transitives from 15 academic Englishes in Africa, America, Asia, Europe and the Pacific. AUTHORS (forthcoming) zoom in on the role that authorial presence plays in the choice of active over passive in six ENL varieties. The focus in this study is on language-internal factors predicting voice alternation.

2.2 Factors predicting choice of passive voice

Several factors are known to determine use of passives in general. In this study, we

look at four that lend themselves to modelling in a multivariate analysis, two syntactic, one semantic and one discourse-pragmatic factor.

Complexity of the verb phrase

English is a language that marks certain categories in the VP (voice, aspect, and occasionally also tense) periphrastically. This can give rise to quite complex VPs, such as *they will have been being chased*. While such maximally complex VPs are rare (see Hundt 2013: 167), combinations of the perfect or progressive with passive are used fairly regularly. One of the strategies in second language acquisition (especially by adults) is to reduce structural complexity; but errors may also lead to more complex patterns (see Kortmann & Szmrecsanyi 2009 or Thomason 2013). We would therefore expect to see differences between ENL and ESL varieties in the combination of passive voice with other verbal categories, and a tendency for learners to avoid passives in complex VPs.

Weight

From a syntactic perspective, passives are often seen to rearrange the order of elements so that long, heavy constituents occupy final position, thus conforming to the Principle of End Weight, as first put forward by Behagel (1909, 1930). Since this is likely to be a universal, cognitive processing constraint, we expect constituent weight to play a significant role in the choice between active and passive across ENL and ESL varieties. If anything, the factor might be more marked in the ESL varieties than in BrE and AmE.

Animacy

From a semantic standpoint, English, like many other languages, tends to have human subjects and topics, as encapsulated in the Animacy Hierarchy originally proposed by Silverstein (1976): human > non-human animate > inanimate. Passives have been shown to contravene this hierarchy (Seoane 2009: 375-379). At the same time, animacy as a predictor for syntactic variation has turned out to be subject to regional variation in WEs in various studies that model frequency effects in grammar (e.g. Hinrichs & Szmrecsanyi 2007, Bresnan & Hay 2008 or Bresnan & Ford 2010). It is possible that it affects voice alternation to different degrees across WEs, and that ESL varieties might prefer animate subjects in both actives and passives.

Givenness of the active object and passive subject

In a language with fixed word-order, like English, passives work as a possible order rearrangement strategy to conform to the given-before-new principle, which concerns the degree of accessibility of information to discourse participants. This pragmatic information status (given vs. new) is a scalar concept rather than a dichotomy, and several taxonomies have been proposed to measure the degree of givenness (see Seoane 2012). The accessibility of information depends not only on the linguistic context (what has already been mentioned and therefore activated in the interlocutors' minds) but also on the extra-linguistic context (knowledge of the world, the perception of the immediate extra-linguistic context by interlocutors). Though there is not a one-to-one correlation between givenness and definiteness (definite *the chair* is supposed to convey given meaning vs. *a chair* which is typically used to introduce a new referent in discourse) analyzing the degree of definiteness of active object and passive subject can help identify contrasts between given and new elements.

2.3 Englishes selected

This paper aims to add to previous studies by looking at contextual factors which may play a role in voice alternation. We also compare ENL and ESL varieties, thus adding to a growing body of probabilistic research into WEs. The typical pattern to emerge from such studies is that of a shared core probabilistic grammar with variety-specific peculiarities at a more fine-grained level of analysis. Szmrecsanyi et al. (2016: 133) call this 'probabilistic indigenization', which they define as

> the process whereby stochastic patterns of internal linguistic variation are reshaped by shifting usage frequencies in speakers of post-colonial varieties. To the extent that patterns of variation in a new variety A, e.g. the probability of item x in context y, can be shown to differ from those of the mother variety, we can say that the new pattern represents a novel, if gradient, development in the grammar of A. These patterns need not be consistent or stable ..., but they nonetheless reflect the emergence of a unique, region-specific grammar.

For the present study, we selected BrE and AmE as our ENL reference varieties. In our choice of post-colonial contact varieties, we focus on the Asia-Pacific region, selecting Singapore (SingE), Hong Kong (HKE), Indian (IndE) and Fiji English (FijE) as varieties that are historically related to BrE, and Philippine English (PhilE) as a second-language

variety deriving from AmE.

In his seminal work on post-colonial Englishes, Schneider (2007) outlines five developmental stages for the evolution of new Englishes. Figure 1 shows that the second-language varieties we selected for this case study can be found at stages two/three and four along this cycle:

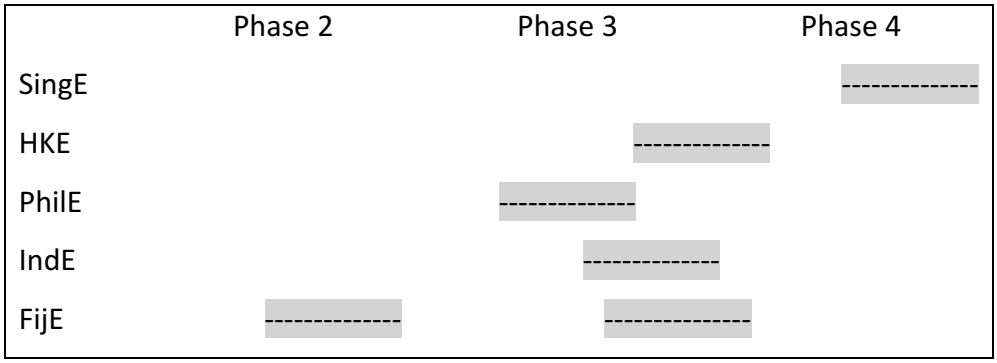

Figure 1: Developmental phase of post-colonial varieties of English according to Schneider's (2007) model

According to Schneider (2007: 114-18), FijE is the least advanced variety here, i.e. it is still at the stage of exo-normative stabilization in a bilingual setting with the original 'homeland' as the norm-giving centre but only incipient structural borrowing; other studies (Geraghty et al. 2006, Zipp 2014, Hundt et al. 2015) indicate that it has progressed to stage three (nativization) and might even be showing the first signs of moving to stage four (endo-normative stabilization). HKE, IndE and PhilE have progressed to the stage of structural nativization, which sees the development of distinctly local structural characteristics; HKE and IndE are moving towards stage four, whereas PhilE is likely to remain at stage three; SingE, finally, has moved on to the stage of endo-normative stabilization, i.e. the emergence of a relatively homogenous local variety that finds acceptance in the speech community (see Schneider 2007).

In addition to differences in the development towards new Englishes in Schneider's model, the five contact varieties were selected because the (dominant) substrate languages potentially allow for different outcomes with respect to structural factors that influence alternation between active and passive.

2.4 Passives and language contact

We use the term 'passive' here rather than 'voice' because passive can be grammatical but also semantic and/or pragmatic. In English, for instance, passive voice is marked

grammatically with catenatives *be* or *get* followed by a past participle.[2] English, too, has constructions which can be argued to be notionally passive but which are formally active in terms of 'voice': examples would be *'*middles' (*e.g. The book reads well)* or remnants of the earlier passival (e.g. *Coffee is now serving*) (see Hundt 2004, 2007).

According to Keenan (1985: 247), a basic passive is characterised by the following properties: "(i) no agent phrase … is present, (ii) the main verb (in its non-passive form) is transitive, and (iii) the main verb expresses an activity, taking agent subjects and patient objects." In other words, basic passives, from a typological point of view, do not mark voice morphologically on the verb. In the following, we look at the main substrate languages for the ESL varieties we investigate to see whether they mark passives morphologically, in other ways, or not at all.

Substrate languages for Singapore and Hong Kong English

The most important substrate languages in these two East Asian countries are Mandarin and Cantonese Chinese, respectively. Other substrate languages relevant for SingE are Malay and Hokkien (and to a lesser extent Tamil, spoken by the Indian population).

In Hokkien (a group of Min Nan Chinese dialects), Mandarin (an isolating language of the Sino-Tibetan group) and Malay (an agglutinative language belonging to the Austronesian family of languages) verbs do not inflect for voice. In Mandarin, for example, *bei* has grammaticalised as a "function word with no inherent meaning other than passiveness marking" (Xiao et al. 2006: 125); it can appear in both long and short passives (i.e. those with or without an explicit agent) without any change in the verb form. Other less frequent passive markers in Mandarin are *gei*, *jiao* and *rang* (Xiao et al. 2006: 125). As for the frequency and distribution of the passive in Mandarin, Xiao et al. (2006: 124-141) observe that passives are less frequent in Mandarin than in English because they normally convey negative and adversative meanings; only recently, because of Western influence, have passives started to be used with positive or neutral meanings (Xiao et al. 2006: 135, Gunn 2017). Passives are especially infrequent in academic writing since, unlike in English, they do not mark objectivity and impersonality.

[2] Only *be* but not *get* also satisfies the criteria for auxiliarihood.

The passive in Cantonese is similar to that of Mandarin, the main difference being that the passive marker *bei* can only appear in long passives in Cantonese (Matthews and Yip 1994: 149), i.e. passives with an overtly expressed agent. As is the case in Mandarin, passives in Cantonese are less frequent than in English, because there are other strategies in the language which readily topicalise objects and because they are still strongly associated with the expression of adversative meaning (Matthews and Yip 1994: 150).

Substrate languages for Philippine English

The linguistic situation of the Philippines is one of intense multilingualism (McFarland 2008: 143). The most widely spoken indigenous language is Tagalog (and the standard version Filipino since 1987), which – like Malay – is an agglutinative language of the Austronesian family. The voice system in Tagalog is a very controversial topic, especially as far as the status of the grammatical subject and the typological nature of the language are concerned (see Shibatani 1988: 85-142). In brief, Tagalog has a goal-topic construction, in which the verb is marked inflectionally as requiring the patient/goal of the action (and not the agent) to be the topic. In addition, it sometimes takes an actor complement, the semantic equivalent of an English *by*-phrase (Schachter & Otanes 1972: 73; Shibatani 1988: 86-89).

Dominant substrate influence on Indian English: Hindi

In India, too, we are looking at a multilingual situation with various local languages that are likely to have an impact on the English spoken. At the same time, most studies considering substrate influence in IndE typically focus on Hindi, since it is an important indigenous national language. Hindi is a morphologically polysynthetic language of the Indo-Aryan language family. In Hindi, passive VPs consist of the passive auxiliary *ja* following an inflected past participle of the main verb. Most of the passives in Hindi are agentless, their main function being that of eliding an irrelevant agent (Sandahl 2000: 101; Kachru 2006: 93).

Substrate and Fiji English

In the Fiji Islands, dialects of Fijian and a non-standard variety of Hindi (Fiji Hindi) comprise the main substrate influence for FijE. We only comment on Fijian here, as passive marking in the local variety of Hindi is similar to what we find in standard Hindi.

Fijian is an Austronesian (Malayo-Polynesian) language that uses some inflectional endings on the verb and is thus somewhat less isolating than other languages in the South Pacific (Lynch 1998: 130f.). Interestingly, transitivity is one of the categories that are marked with the help of a suffix Fijian verbs (ibid.: 140). There are three broad types of transitives in Fijian: active transitive, passive transitive and transitive only. In order to express passive meaning, Fijian relies on the so-called passive transitive verbs used without the transitive inflectional marker (Geraghty 2008: 28-29). According to Biewer (2009: 366), Fijian employs several inflectional strategies to mark the passive on the verb, and according to Schütz (2014: 104) one of them is to add a stative marker to an active verb, which then has the goal as its subject, but without being able to add the agent in this construction:

> … there is not simple answer to the question: Does Fijian have a passive construction? If "passive" means simply "goal-focussed", the answer is yes. If one insists that a passive sentence contain reference to the actor, the answer is no.

To sum up, all substrate languages have means of expressing the passive notionally, but formal means of doing so are present in the substrate of only four out of the five contact varieties we study. Table 1 summarises the means that are used in the various substrate languages by country.

	periphrastic	inflectional	other
Singapore	-	-	+
Hong Kong	-	-	+
India	+	-	-
Philippines	-	-	-
Fiji	-	+	-

Table 1: Synopsis of grammatical passives in substrate languages

If substrate influence were to play a significant role, one possibility is that the combination of passive with other verbal categories is avoided in SingE and HKE, where the dominant substrate languages mark passive outside the VP and where complex VPs may occur less frequently than in other WEs. We further expect 'givenness' to play a more important role in PhilE than in the other ESL varieties if substrate plays a role: according to Maratsos (1988: 133)

> [e]ach sentence in Tagalog has to give at least one argument that is pragmatically given. If the sentence has a verb, the verb must have an agreement marker that marks the grammatical case (nominative, accusative, dative, or benefactive, typically) of this given argument.

These are possible outcomes of more subtle, probabilistic substrate influence than the one evidenced in structural transfer that gives rise to constructions like the *kena*-passive in SingE.

2.5 Contact-induced phenomena and the passive

In addition to substrate influence, processes of language acquisition may give rise to features in the contact varieties which, in turn, may affect the realization of the *be*-passive in the ESL varieties investigated here. A wide-spread phenomenon is final consonant cluster reduction, for instance, which in turn may lead to the use of unmarked past participles, as in the following example (retrieved from the *Global Web-based English* corpus):[3]

(1) … the land *was consider* worthless.
(http://www.phuket101.net/2011/06/bang-tao-beach.html)

This is said to be especially frequent in SingE and HKE, as expected from the isolating nature of the substrate languages and from the fact that their substrate languages tend to simplify consonant clusters (Deterding 2007: 17-19; Deterding *et al*. 2008). Unmarked participles are likely to pose a problem for the automatic retrieval of passives from syntactically annotated data (see section three).

Auxiliary deletion is another common feature found in various contact varieties of English. Biewer (2015: 186) provides the following passive with auxiliary deletion from her corpus of FijE (interview data):

(2) and in school when we speak Fijian we are we Ø always punished/
(Fiji/WI.txt)

This kind of feature is much less likely to occur in published academic writing than in spoken interaction, however.

Previous research has also shown that some L2 learners of English tend to over-passivise, that is, they extend the use of the passive to verbs that do not allow for a passive in ENL varieties; these include unaccusative verbs that may or may not have transitive counterparts in some contexts (*Tom ate / Tom arrived*) and unergatives,

[3] GloWbE can be accessed at the following site http://corpus.byu.edu/glowbe/; for background information and a critical discussion of the advantages and shortcomings of this resource, see Davies and Fuchs (2015).

such as *Tom laughed* (Kondo 2005: 129). An example of overpassivization from learner English is the following (from Kondo 2005: 156):

(3) The coin *was vanished* instantly.

On the whole, the structurally markedly different passive constructions that are due to general processes of second language acquisition are quite likely to have been edited out in published academic writing. Hundt et al. (2016) test for recall of automatically retrieved transitive active and passive constructions and are able to show that such characteristically ESL constructions are very infrequent in the academic part of the ICE corpora.

2.6 Research questions

On the basis of previous research and the choice of varieties we investigate, we can formulate the following research questions:

- Will we see a divide into first- and second-language varieties when it comes to the internal factors predicting the choice of a passive over an active?
- More specifically, will we find regional differences in the role that animacy of the subject plays as a predictor?
- For the second-language varieties, can we observe possible influence of substrate languages or the process of second-language acquisition in the effect size that the factors have in predicting voice alternation?

3. Data and methodology

3.1 Parsed ICE corpora

Our data come from the published academic writing section of the ICE corpora (see Table 2). These comprise ten 2,000-word samples each from the humanities, social sciences, natural sciences and technology, and thus a total of approximately 80,000 words per variety and 560,000 words for all Englishes investigated in this paper. For a study of a frequent alternation as that between active and passive, such a relatively small corpus yields ample evidence.

ICE-GB	Great Britain
ICE-US	United States of America
ICE-SIN	Singapore
ICE-HK	Hong Kong
ICE-IND	India
ICE-PHI	The Philippines
ICE-FJ	Fiji

Table 2: List of ICE components (abbreviations) included in the study

The ICE corpora were syntactically annotated with the probabilistic dependency parser (Pro3Gres; Schneider 2008). The annotation process includes part-of-speech tagging, chunking and parsing and makes automatic retrieval of passives and active transitives possible (see next section).

3.2 Data retrieval

Hundt et al. (2016) comment in detail on the automatic retrieval of central *be*-passives and active transitives. We make use of the same approach to retrieve data from the seven ICE components. We then coded sets of 200 randomly sampled instances per variety, 100 passives and actives, each, and equal numbers of actives and passives per subdiscipline. We manually excluded false positives such as the following from our data.

(4) … more Fijian businesses would *bite* dust if Government did not act now (ICE-FJ W2A 011)

(5) Therefore it can be assumed that those who *enter* a university have chosen to do so … (ICE-SL W2A 004)

(6) One of the problems I *wish to address* is the degree to which Frankish uncial in the late eighth and the ninth centuries is indeed artificial rather than natural. (ICE-GB W2A 008)

Likewise, only those central *be*-passives were included in our investigation that had active transitive counterparts, excluding for instance adjectival passives as in (7):

(7) Mesozoic isolated platforms *are* well *represented* in the Tethyan region. (ICE-GB W2A 023)

Having excluded non-relevant material from our randomly sampled concordances, we initially ended up with a total of *N*=1285 variants (roughly equal amounts of actives and passives) that we coded for various internal factors.

3.3 Factors coded for

Complexity of the verb phrase

We coded the complexity of the VP in terms of TAM combinations with voice, distinguishing between simple present and past, progressive present and past, present and past perfect as well as combinations of central modals with active or passive. As it turned out, the factor was only weakly significant as long as individual patterns were coded separately (e.g. present and past progressive). We therefore decided to group complex VPs (perfect, progressive and modal) together and contrast these with simple (both present and past) VPs.

Weight

The concept of weight has been defined as number of syllables, words, nodes and phrasal nodes (Rosenbach 2007; Hawkins 2004). As Wasow & Arnold (2003) point out, however, it is still not clear whether the different ways of measuring constituency weight yield different results (see Seoane 2009: 370-371). For the purposes of our study, we measure constituent length in number of words; for constituents that are longer than four words, we use brackets (4-9 coded as '4', 5-10 as '5', etc.). This means that in some examples, the subject may actually be slightly longer in total number of words than the object, even though in our analysis they show up as having the same length in terms of the bracket they occur in. An example is given in (8), where the subject amounts to 9 words whereas the object comprises only 5 words, with both falling into our coding bracket '4', i.e. constituents between 4 and 9 words in length:

(8) … [the high amplitude specular signal from the 0 probe]$_S$ emphasises [the surface of the defect]$_O$. (ICE-GB W2A 031)

Also, we counted the number of words of the syntactic unit functioning as the subject and object/agent rather than the semantic subject. In the following example (9), it is the relative pronoun that is the syntactic subject (1 word) of the active transitive verb *change,* even though its antecedent, and thus the semantic subject (*massive upheavals* …), is much longer.

(9) World War II produced massive upheavals beyond the preemptive political and military events <u>that</u> *changed* the future course of history. (ICE-US W2A 013)

Animacy/Semantics

The factor 'animacy' is not a binary one but a gradient (see section 2.2).[4] Nonetheless, we coded for animate subjects/objects and inanimate ones, adding a third category (unclear) only if animacy could not be determined on the basis of the context.[5] This means that the factor 'animate' in our analysis includes a fairly broad range of semantic concepts, such as prototypical animate and human NPs, e.g. personal pronouns as in (10) and the authors' names in (11), but also collective nouns such as *government* and personifications such as *Rome* in examples (12) and (13), respectively.

(10) what *I* found was that almost all his poetry is filled with expressions of his need for and love of other people. (ICE-GB W2A 003)
(11) *Wilson & Henderson* (1966) described U. ambiguus and P. allii as the two species in the U.K. with P. porri and P. mixta (on chives) as synonyms of P. allii (ICE-GB W2A 028)
(12) The Griffiths Report was not received with great enthusiasm by *the Conservative government …* (ICE-GB W2A 013)
(13) … and in the stress of war *Rome* conceded what they had sought. (ICE-GB W2A 001)

Examples of constituents whose semantics are unclear are given in (14) and (15):

(14) These steps are iterated until an acceptable solution is reached. (ICE-GB W2a 016)
(15) So that neither *the user or system* is overwhelmed by large result sets, the size of result sets is limited to 100 items … (ICE-US W2A 038)

The agent could be an abstract process in example (14) or a human agent, whereas the combination of an animate with a clearly inanimate NP in (15) renders the subject constituent neutral, so to speak, with respect to the factor 'animacy'. Cases of unclear animacy coding are discarded in the subsequent analyses.

In short passives, where the semantics of the agent could be inferred from the context, animacy was coded for as outlined above. In example (16), for instance, the inferred object/agent is coded as 'animate', while in example (17), the inferred agent of the passive phrase is coded as 'inanimate'.

[4] Zaenen et al. (2004), for instance, distinguish eleven categories, including two for instances where the annotator was unsure of the coding. This indicates that there is no universal solution to the problem of coding for the factor 'animacy'.

[5] These had to be excluded in the mixed-effects modelling because of model convergence issues, leaving us with 1,168 items.

(16) For example, thermotropic responses of seedlings *have* occasionally *been noted* (Aletsee 1962a), […]. (ICE-GB W2A 025)

(17) Leaching is a continuous process in the dunes because the sands do not overlie a bedrock *which is being gradually weathered*, as do most terrestrial soils, but if anything are accreting more depth of sand. (ICE-GB W2A 022)

Definiteness/Givenness

Finally, the factor 'givenness' is not really a binary one, either (see section 2.2). There are various ways of coding for 'givenness' (for a discussion, see Dreschler 2015: 83-86). We decided to approximate this factor by coding both active objects and passive subjects as either 'pronominal', 'definite', or 'other'. Due to data sparseness, pronominal and definite constituents were then conflated under the label 'definite', all other instances were coded as 'indefinite'. Among definite noun phrases, we also included instances with proper names:

(18) *Mrs. Roca* … expressed her frustration over the intricate process she had to go through to face her husband about the separation. (ICE-PHIL, W2A 015)

3.4 Statistical modelling

In order to tease apart influence of the predictors on voice alternation across the seven varieties of English, we make use of two multivariate techniques. First, we model the data using a tree and forest approach as first advocated by Tagliamonte & Baayen (2012). Both are variants of permutation testing that do not assume a certain distribution of the data but build a model by resampling from the input. The random forest analysis provides information on overall variable importance and the single conditional inference tree (ctree) allows us to tap into and visualise possible interactions between predictor variables.[6] The ctree splits the data recursively into smaller subsets according to those predictors that co-vary most strongly with the outcome. For each binary split, the data is inspected for the predictor that best preserves the homogeneity of each split (e.g. all actives versus all passives) at the customary significance level of $\alpha = 0.05$. The splitting process is then repeated until no further splits can significantly increase the subsets' homogeneity with regard to the

[6] We used the party package (Strobl, Hothorn and Zeileis 2009) and the partykit (Hothorn et al. 2006) as implemented in R, respectively, to fit the conditional random forest and the ctree.

outcome variable. The bottom-most barplots provide information on the observed proportions of outcome variants in that particular split. Single trees have the disadvantage that they very much hinge on the dataset at hand and are hence subject to a high degree of variability. Conditional random forests resample over a predefined number of trees using a conditional permutation scheme and are thus particularly robust to, for instance, predictor collinearity. In addition to predictor rankings obtained through the random forest, and in order to gain a more robust insight into the predictors' variability across our set of geographically dispersed varieties of English, our second approach uses mixed-effects modelling.[7] This allows us to evaluate the significance of individual predictor variables in the voice alternation (Hosmer and Lemeshow 2000; Pinheiro and Bates 2000). A mixed-effects model makes adjustments to the model's predictions from the fixed effects by including random effects in the modelling process. Random effects account for idiosyncractic variation by group that is specific to the dataset, such as lexical items or texts sampled. By using mixed-effects modelling, we are able to generalise beyond the particular dataset at hand to, for instance, all texts of academic English. In particular, it allows us to look into possible interaction with the predictor 'variety', which our sampling precluded from emerging as an important effect in the random forest analysis.

4. Results

4.1 Descriptive statistics

Complexity of the verb phrase

The proportional distributions in Figure 2 do not confirm our hypothesis of substrate influence, which would have predicted that passives are avoided in complex VPs in varieties where the main substrate marks voice outside the VP (i.e. SingE and HKE). Instead, Figure 2 shows that passives are proportionally preferred over actives in complex VPs in academic writing, with the exception of FijE. In simple VPs, actives tend to be preferred over passives, with the exception of FijE and HKE, which prefer the passive, and PhilE, which shows an even distribution.

[7] We make use of the glmer() function from the lme4 package in R (Bates et al. 2015) for this.

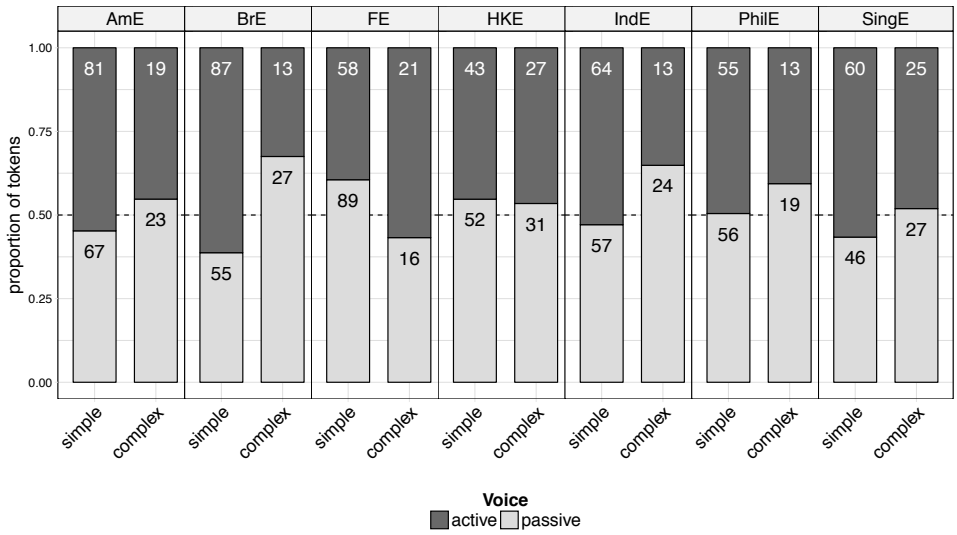

Figure 2. Proportional distribution of simple and complex VPs by voice variant. Raw numbers provided in each bar.

Weight

Figure 3 shows the smoothed conditional means of the proportion of passives (y-axis) by increasing length (x-axis) of the subject (solid line) and the object (dashed line): the proportion of passives increases with increase in subject-length and the proportion of actives increases with increase in object-length, in accordance with the principle of end-weight.

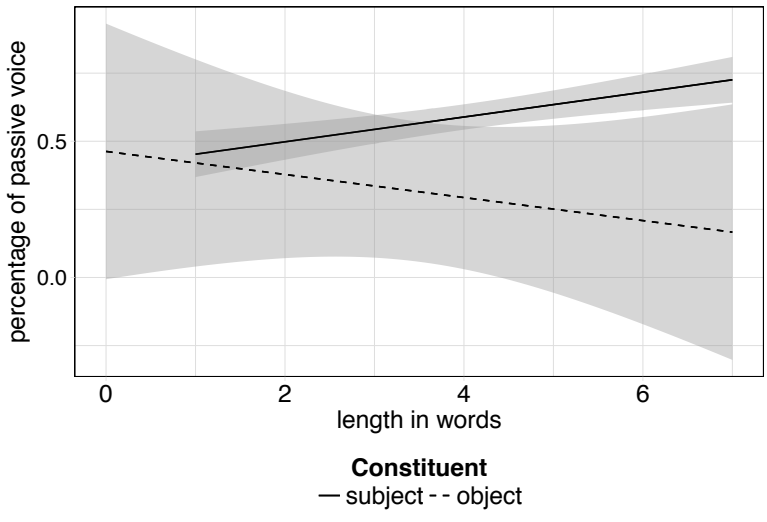

Figure 3. Smoothed conditional means of the proportion of passive variants by increasing object length (dashed line) and subject length (solid line). Note that objects can have 0 length when they are not present in the construction.

Animacy/Semantics

Regarding the proportional distribution of SemanticsOfSubject, actives prefer animate

subjects whereas inanimate subjects dominate in the passive (see Figure 4).

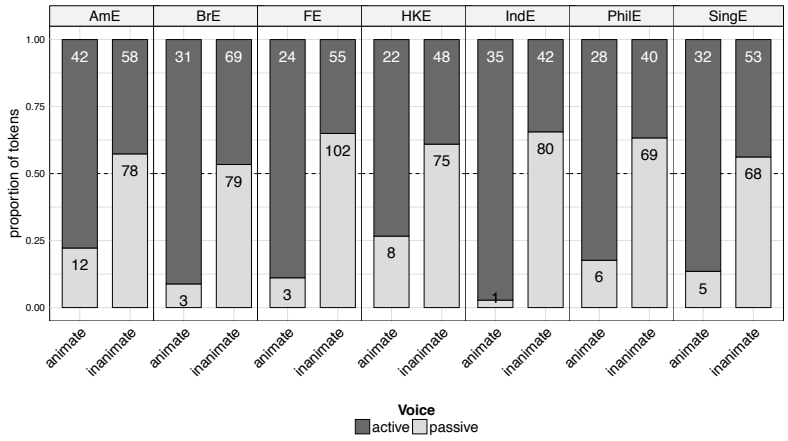

Figure 4. Proportional distribution of animate and inanimate subjects by active and passive voice variants.

With respect to SemanticsOfObject, the situation is reversed with actives preferring inanimate objects and passives showing an overall preference for animate objects (see Figure 5).

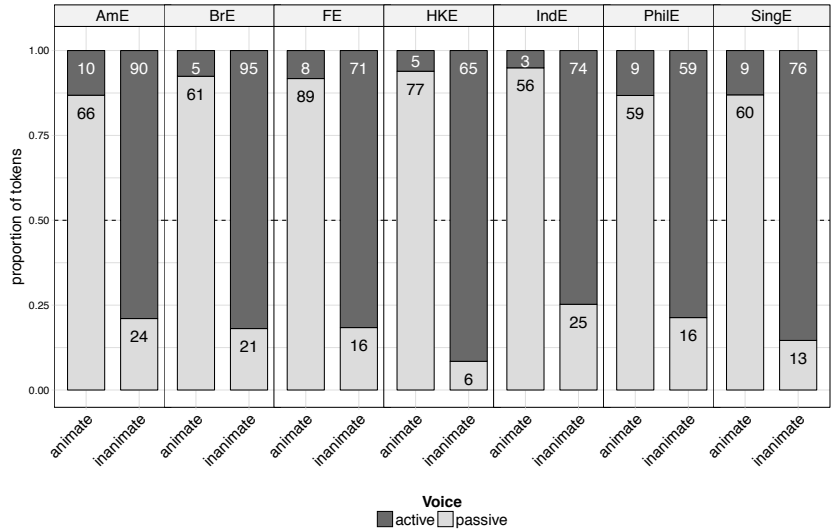

Figure 5. Proportional distribution of animate and inanimate objects by active and passive voice variants.

Definiteness/Givenness

The general trend across varieties is for definite patients to be more likely in passives, where they function as subjects, than in the active voice. Indefinite patients, on the other hand, are more likely in actives, where they function as objects. Definiteness/Givenness has thus the expected effect reported in the literature.

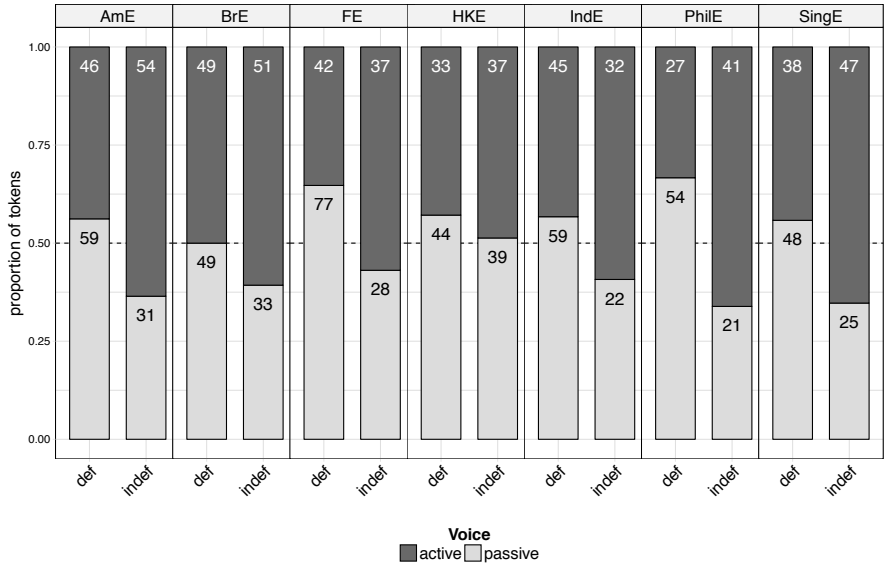

Figure 6. Proportional distribution of definite and indefinite patients by active and passive voice variants.

4.2 Forest and tree analysis

Figure 7 ranks the importance of the predictor variables. It shows that length of object (related to the principle of end-weight) is the most important predictor for the choice of a passive over an active VP, followed by animacy (notably the semantics of the object). All other factors are less important.

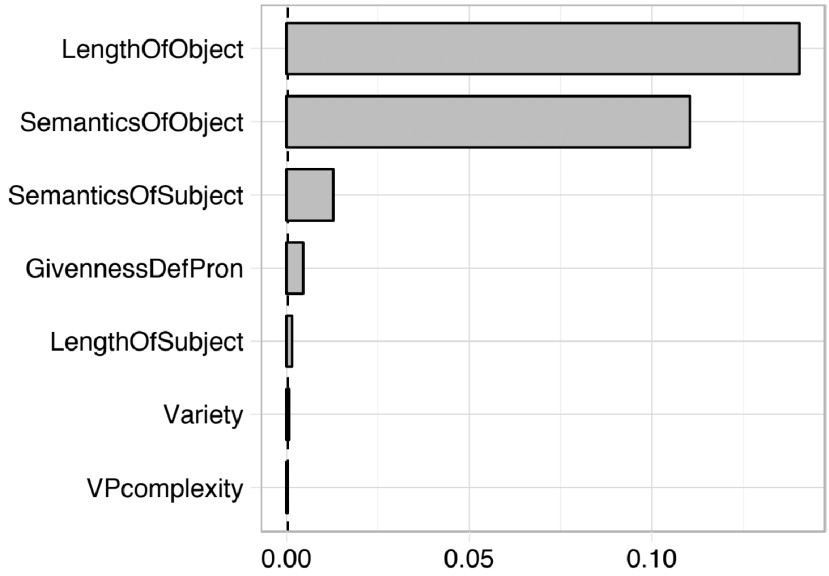

Figure 7. Overall variable importance (random forest analysis) for voice alternation in academic English[8]

[8] We set mtry=3 and ntree=2000, following recommendations by Strobl (p.c.). Somers2 Dxy returns a prediction accuracy of 0.951 and a *C*-index of 0.975, which is above the level of 0.8 recommended e.g. in Tagliamonte & Baayen (2012: 156).

The output of a conditional inference tree is shown in Figure 8.[9]

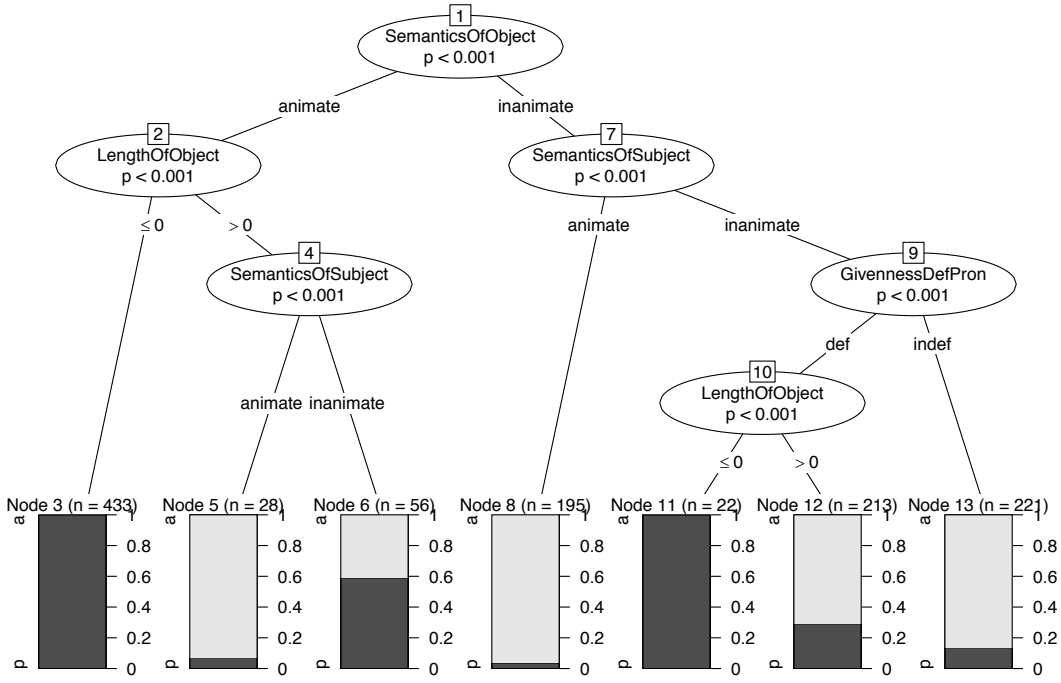

Figure 8. Conditional inference tree of passive (p) over active (a) choice in academic English.

The ctree returns SEMANTICSOFOBJECT as the most predictive factor (Node 1). Inanimate objects are more likely to be realised in the active variant, while animate objects (and unclear cases) favour the passive. Proceeding down the left side of the tree, we see that among animate objects and unclear cases, the passive variant is categorically chosen if the object is inexistent (Node 2). This is not surprising given that animate agents are hardly ever expressed in passives and often inferred from the context (see example (16)). In cases where the object/agent is present, inanimate subjects have a stronger preference for the passive than animate subjects (Node 4).

Moving down the right side of the tree where the object is inanimate (from Node 1), we see again a split by semantics of the subject (Node 7): In the case of inanimate objects, the active variant is preferred if the subject is animate (in other words, both subject and object are expressed in active transitive clauses) (Node 8).

[9] The stability of the classification was confirmed through a second tree fitted with a different random seed. Classification accuracy of the conditional inference tree is 89.4%, i.e. significantly above the baseline of 50.4%; the *C*-statistic (0.894) is also above the recommended level of 0.8. Note that only splits with a *p*-value of 0.05 or lower were allowed when building the tree.

When both subject and object are inanimate, the definiteness of the syntactic object in the active and the syntactic subject in the passive clause as well as length of the object become relevant (Nodes 9 and 10). Passive is the categorical choice if the syntactic subject is definite or pronominal and the object/*by*-agent is not explicitly expressed (passive is the only viable option in those cases) (Node 11). If the object is expressed, however, actives are the preferred option (Node 12). If both object and subject are inanimate the active is preferred with indefinite subjects (Node 13). Neither 'variety' nor 'subdiscipline' show up as significant factors most probably due to the sampling process.

4.3 Mixed-effects model

Our initial model included all factors (Section 3.3) as fixed effects in interaction with VARIETY (no other higher order interactions were considered). The numeric factors LENGTHOFOBJECT and LENGTHOFSUBJECT were standardised by two standard deviations and centered around the mean in order to reduce potential covariation among the numeric and other predictors and to create a predictor with an effect size on a scale comparable to that of a binary predictor (see Gelman 2008). VARIETY was coded using sum contrasts whereby the proportion of responses for each level is compared against the grand mean across all levels (see Menard 2010: 97). Random effects included a simple random intercept for File in order to account for idiosyncracies in the texts sampled to represent academic writing.

Model selection followed the backwards elimination procedure outlined by Zuur et al. (2009: 120–122). Because of convergence issues with the full set of interactions, we additionally fitted a fixed-effect model in order to identify interactions that did not significantly improve model fit and cross-tabulated the data to find sparse cells. The exclusion of the most negligible predictor (SEMANTICSOFSUBJECT) from any interactions made model convergence possible. The predicted outcome of the model was the log odds of the passive variant. Next, we identified and excluded any other interaction terms that did not significantly improve model fit.

The final model (19) includes a by-file random intercept, as well as an interaction of VARIETY and SEMANTICSOFOBJECT, GIVENNESS, and LENGTHOFSUBJECT. None of the other initial interactions with VARIETY turned out to be significant.

(19) Passive model 1; Response = {active, passive}
Response ~ (1|FILEID) + SEMANTICSOFSUBJECT + VPCOMPLEXITY + LENGTHOFOBJECT + VARIETY * (SEMANTICSOFOBJECT + GIVENNESS + LENGTHOFSUBJECT)

Classification accuracy of the model is 88.5% which is significantly better than the baseline of 50.43% when always choosing the most frequent (passive) variant (p_{binom} < 0.001). Summary statistics further give a very good index of concordance C = 0.946, which indicates a good model fit (see Baayen 2008). The condition index κ=10.9 indicates existent but not harmful collinearity (Baayen 2008: 182). The variance inflation factor (VIF) for each of the factors points out that much of the estimated variance of all higher order interactions with VARIETY is associated with the corresponding main effect and with other interaction effects with VARIETY. We will therefore exercise extra caution when interpreting these results.

To evaluate the model, we randomly divided our dataset 100 times into a training and a test set. Next, we fitted the model to each training set and calculated the predictions for the corresponding test set. The accuracy and index of concordance C of each model was then measured and compared to the original model. Mean accuracy of the 100 models was 86.8% which corresponds to a drop of 1.7% from the original model. *C*-statistic was an excellent 0.946.

Main effects

The coefficients of the main factors in the model are summarised in Table 3.

Factor	**β**	**SE**	***p***
(INTERCEPT)	-4.25548	0.45754	<0.001 ***
SEMANTICSOFSUBJECT			
animate ⇨ inanimate	2.64365	0.31323	<0.001 ***
SEMANTICSOFOBJECT			
inanimate ⇨ animate	3.31353	0.38393	<0.001 ***
GIVENNESS/DEFINITENESS			
indef ⇨ def	1.10909	0.22423	<0.001 ***
LENGTHOFOBJECT	-2.22938	0.28321	<0.001 ***
LENGTHOFSUBJECT	0.91655	0.29722	<0.00204 **

Table 3: Main effects of individual factors in the model. Predictions are for passive voice. Only significant factors shown.

The column labelled β indicates the estimates of the coefficients on a logit-scale. Positive values signal a preference for passive (the predicted outcome), negative values a preference for the active voice. SE specifies standard errors. The results of

the logistic regression analysis can be summarised as follows:

First, the main effects in the model have largely the predicted influence on the choice of voice variant given previous literature. The probability of the passive increases if the subject is inanimate instead of animate, if the object is animate instead of inanimate, if the VP is simple instead of complex (not a significant factor) and with every unit increase in the length of the subject. In other words, passives tend to be used with an inanimate long subject and animate object/agent. The likelihood of a passive also increases if the subject of the passive voice or the object of the active voice is definite, that is, definite constituents (e.g. pronouns) are more likely to be used in subject position in passives than in object position in active voice. This follows naturally from the ordering preference in the human processing system in that speakers tend to use definite/given constituents before new information (see Wasow 2002: 30 and literature cited therein; also Ransom 1979). The probability of a passive further decreases with every unit increase in the length of the syntactic object, i.e. the agent in the passive voice. Table 3 only reports significant factors; in other words, VP complexity turns out not to be a significant predictor. Cross-varietal differences do not emerge as a main effect in the overall choice of passive vs. active because equal numbers of passives and actives was sampled per variety and subdiscipline. The next section therefore specifically looks at possible interaction effects with VARIETY.

Interaction terms

Table 4 reports all significant interaction terms between VARIETY and the language-internal factors SEMANTICSOFOBJECT, GIVENNESS and LENGTHOFSUBJECT. Note that none of the other interaction terms contribute significantly to the model fit. If the coefficient estimates of a main predictor (for instance, SEMANTICSOFOBJECT) and its interaction term (VARIETY : SEMANTICSOFOBJECT) have the same +/- sign, the effect is stronger in that specific variety (compared to all other varieties). If they have opposite signs, the effect of that factor is weaker in that specific variety (or possibly even reversed). Predictions are for the passive variant.

Factor	β	SE	*p*
VARIETY : SEMANTICSOFOBJECT			
HKE + animate	3.37523	1.49674	0.02414 *
VARIETY : GIVENNESS/DEFINITENESS			
HKE + def	-1.48428	0.68535	0.03033 *
PhilE + def	1.12773	0.54831	0.03971 *
VARIETY : LENGTHOFSUBJECT			
HKE	3.38519	1.40018	0.01562 *

Table 4: Significant interaction effects between Variety and language-internal factors in the model. Predictions are for passive voice.

Zooming in, the interaction effects in the model indicate that academics from Hong Kong deviate from the global average with regard to the effect of all three language-internal predictors, and those from the Philippines with regard to only one.

- Animate objects are more likely to be used in passives in HKE academic writing; the effect of SEMANTICSOFOBJECT is thus stronger in that variety (see Figure 3).
- At the same time, the effect of GIVENNESS/DEFINITENESS is weaker in HKE, that is, if the constituent is definite, the likelihood of passive voice is not as strong as in the other varieties (see Figure 4).
- The effect of GIVENNESS/DEFINITENESS is stronger in the academic writing of researchers from the Philippines, i.e. they are more likely to use passive voice if the constituent is definite (see Figure 4).
- The effect of LENGTHOFSUBJECT is stronger in academic writing in HKE compared to the global average: academics from Hong Kong are more likely than academics elsewhere to use a passive when the syntactic subject increases in length (see Figure 5).

Figures 9 to 11 report the probability of passive voice given the predictor's level, for instance animate vs. inanimate (y-axis), across the seven varieties (x-axis). Fitted effects were plotted using the effects package in R (Fox 2003).

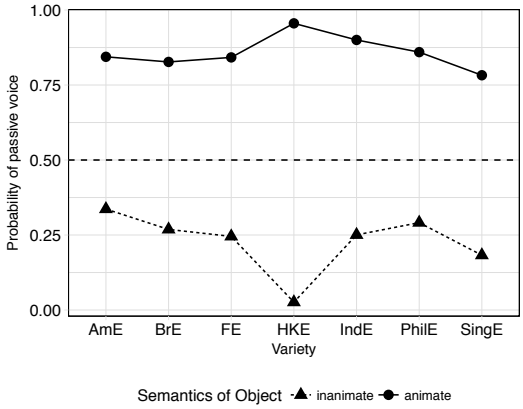

Figure 9: Passive voice and Semantics of object

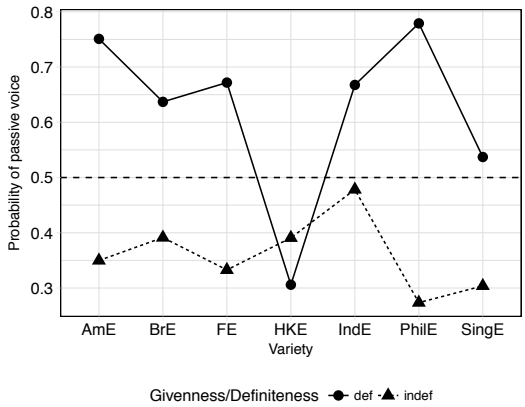

Figure 10: Passive voice and Givenness

Figure 9 shows that the passive is more likely with animate objects and active voice with inanimate ones. Moreover, the difference in the effect is strongest in HKE. The graphs in Figure 10 reveal that the probability of passive is generally higher with definite than indefinite constituents: Givenness/Definiteness has a reverse effect in HKE and a stronger effect in PhilE academic writing. Finally, with respect to the effect that the length of the subject has on the choice of a passive over an active VP, only academic texts from Hong Kong show a significant effect of this predictor, but none that could be easily explained by substrate influence (Figure 11).

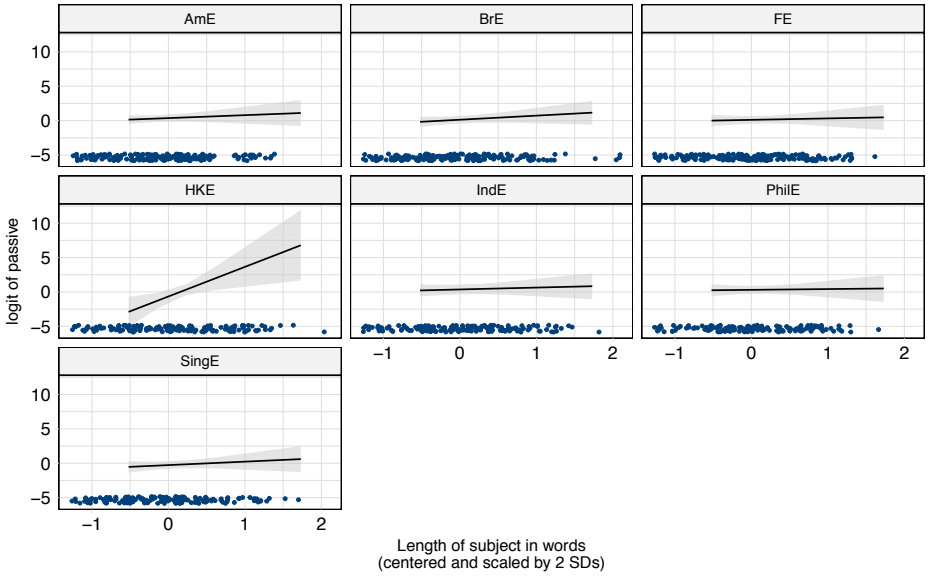

Figure 11: Effect of increase in subject length by variety.

5. Summary and discussion

The analyses show that the choice of a passive over an active VP in academic writing is only partially subject to regional variation. Importantly, there is no straightforward difference between ENL and ESL varieties (i.e. types of Englishes) when it comes to the factors predicting choice of a passive. It is only in academic writing from Hong Kong and the Philippines that we can observe very subtle cross-lectal variability which is to be found in the effect size of LENGTHOFSUBJECT, SEMANTICSOFOBJECT and GIVENNESS. Overall then, the use of passive voice in academic writing is fairly homogenous across different varieties of English, be they norm-providing ENL varieties or ESL varieties. How do these results connect with our initial research questions?

Previous research into other morphosyntactic alternations found that 'animacy' was often a factor that had variable influence across national varieties (see section 2.2). 'Animacy' is relevant for voice alternation in two (related) respects. Firstly, the passive is a construction that allows the subject to be inanimate, contrary to the usual animacy hierarchy.[10] Secondly, the construction also allows the demotion of the agent (typically animate) to the object relation (either overtly expressed as a *by*-agent or to be inferred from the context). While this is a global constraint that holds in academic writing across the varieties investigated, our data show that animacy is a particularly strong factor in Hong Kong academic writing, which prefers passives with animate (underlying) agents. Representative examples that imply an animate, human agent which, however, typically remains unexpressed, are the following:

(20) *It is believed* that the internal cohesion within the relatively self-contained or isolated construction activities will be impaired if adversarial proceedings are employed to settle disputes. (ICE-HK, W2A-014)

(21) Cases of amoebic meningoencephalitis due to Naegleria fowleri *have* also *been seen* locally. (ICE-HK, W2A-021)

(22) What *was forgotten* at the moment or what *was repressed* in the past returns to consciousness in dreams, helping men to re-gain the balance of a healthy mind. (ICE-HK, W2A-002).

[10] Note that occasionally, inanimate subjects are used in active transitives even though the patient in object position is animate, as in the following example: *"Section 29* did empower *local authorities* to promote the welfare of persons who are blind, deaf or dumb, … " (ICE-GB W2A 013).

Givenness turned out to have a weaker effect in HKE than the other varieties. The following are typical HKE examples of passives with non-specific subjects:

(23) *Certain types of building components* are required to be tested for their safety and serviceability. (ICE-HK, W2A-036)
(24) *Occasional cases of fungal abscesses* have also been seen. (ICE-HK, W2A-021)
(25) *Other arts of Chinese* were also combined to the lantern design, … (ICE-HK, W2A-008)

The reverse holds for PhilE, where passivised subjects frequently contain a demonstrative pronoun, a clear marker of 'givenness', or are otherwise clearly 'given' in the preceding context (over and beyond being 'definite'):

(26) *These cut portions* are left for sometime at ambient condition to cure wounds naturally. (ICE-PHI, W2A-033)
(27) At the very outset, *this leitmotif* has to be grasped well because many concepts of Asian philosophy cannot be understood without it. (ICE-PHI, W2A-009)
(28) It is true that an object can require another to perform some task, but, *the actual manipulation of the required data* or service is done in the called object. (ICE-PHI, W2A-038)

That givenness turns out to play such an important role in PhilE academic writing fits in well with the dominant substrate, Tagalog (see section 2.4).

With respect to the factor 'weight', the default in English (and other languages) is to have short subjects and long objects. A good example of a prototypical combination of short subject and heavy object is given in (29), where the reverse achieved by a passive would be highly unlikely.

(29) I began the process of questioning and reexamining the possibility of comparatively studying a select number of successful nonprofit organizations and their leadership to determine what strengths would be revealed regarding their success. (ICE-US, W2A-011)

If both subject and object/agent are animate, length is the deciding factor:

(30) Now in her twilight years, she was a nervous recluse, living off the charity of family and friends, and *she was only visited regularly by a hired reader*. (ICE-US, W2A-007)

In passives, long subjects typically occur without agent *by*-phrases, for instance in contexts where the agent is unknown.

(31) *The two earliest, the Aberdeen fragment written in rustic capitals and the St Gall Vulgate in half-uncials,* may have been written at a time when the hierarchy of appropriate scripts was being worked out. (ICE-GB, W2A-008)

Note that the length of the subject in (31) is actually due to appositions that follow the subject NP. Clausal expansion of the subject NP also contributes the bulk of material in the two extremely 'weighty' passive subjects in the HKE data (examples (32) and (33)), but long, internally complex phrasal NPs are also attested (34):

(32) *Traditionally, site control tests and sampling, such as slump and compacting factor tests to CS1; flow of fresh concrete to BS 1881: Part 105; and making of concrete cubes to CS1, etc,* were performed by the contractor's site staff. (ICE-HK, W2A-036)

(33) *Well-designed population-based studies that aim to access the long-term significance of these new recommendation to diagnosis categories of glucose intolerance in Asian populations* are now needed. (ICE-HK, W2A-028)

(34) *The angiotensinogen M235T (TT) genotype and its possible interaction with the angiotensin converting enzyme deletion/insertion polymorphisms (DI/DD)* have also been reported in Chinese diabetic patients who have increased albuminuria. (ICE-HK, W2A-024)

That long passive subjects should occur with a greater-than-average frequency (see Figure 11) does not find an easy language-internal explanation, however. On the contrary, if substrate were to play a role, we would expect to find exactly the opposite, i.e. relatively short sentences and phrases and no post-modifying clauses within NPs. Our data come exclusively from academic texts, though, and the explanation for the high incidence of long passive subjects in academic writing from Hong Kong might well have to be attributed to the conscious attempt at emulating a western academic writing style resulting in an over-use of constructions that are avoided by researchers with English as their first language (see also Gunn 2017 on stylistic effects in Chinese writing as a result of translation practices).

Finally, our study did not produce a regional difference in the effect size for the factor 'complexity of the verb phrase'. This means that, contrary to our hypothesis, there is no statistically significant trend in ESL academic writing to avoid complex passive VPs, any more than would also be the case in writing produced by academics with English as their first language.

Acknowledgements

TEXT

References

Corpora

ICE *International Corpus of English*

Bibliography

AUTHORS. forthcoming.

Baayen, Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Bao, Zhiming and Lionel Wee. 1999. The passive in Singapore English. *World Englishes* 18(1): 1-11.

Bates, Douglas, Martin Mächler, Benjamin Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1): 1–48.

Behagel, Otto. 1909. Beziehungen Zwischen Umfang Und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25: 110–142.

Behagel, Otto. 1930. Von deutscher Wortstellung. *Zeitschrift für Deutschkunde* 44: 81–89.

Biber, Douglas and Edward Finegan. 1989. Drift and evolution of English style: a history of three genres. *Language* 65: 487–517.

Biewer, Carolin. 2009. Passive constructions in Fiji English: A corpus-based study. In Andreas H. Jucker, Daniel Schreier and Marianne Hundt, eds. *Corpora: Pragmatics and Discourse*. Amsterdam: Rodopi, 361–377.

Biewer, Carolin. 2015. *South Pacific Englishes. A Sociolinguistic and Morphosyntactic Profile of Fiji English, Samoan English and Cook Islands English*. Amsterdam: Benjamins.

Bresnan, Joan and Jennifer Hay. 2008. Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua* 118: 245–259.

Bresnan, Joan and Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English." *Language* 86(1): 168–213.

Davies, Mark and Robert Fuchs. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide* 36(1): 1–28.

Deterding, David. 2007. *Singapore English*. Edinburgh: Edinburgh University Press.

Deterding, David, Jenny Wong and Andy Kirkpatrick. 2008. The pronunciation of Hong Kong English. *English World-Wide* 29(2): 148–175.

Dreschler, Gea. 2015. *Passives and the loss of verb second: A study of syntactic and information-structural factors* (LOT dissertation series 402). Utrecht: LOT.

Fox, John. 2003. Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software* 8(15): 1–27. URL http://www.jstatsoft.org/v08/i15/.

Gelman, Andrew. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine* 27(15): 2865–2873.

Gelman, Andrew and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.
Geraghty, Paul. 2008. *Fijian*. Victoria: Lonely Planet Publications.
Geraghty, Paul, France Mugler and Jan Tent, eds. 2006. *Macquarie Dictionary of English for the Fiji Islands*. Macquarie University: The Macquarie Library.
Gunn, Edward. 2017. Westernization of Chinese Grammar. In Rint Sybesma, Wolfgang Behr, Yuego Gu, Zev Handel and C.-T. James Huang, eds. *Encyclopedia of Chinese Language and Linguistics*, Consulted online on 20 July 2017 <http://dx.doi.org/10.1163/2210-7363_ecll_COM_00000450>
Hawkins, John A. 2004. *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.
Hinrichs, Lars and Benedikt Szmrecsanyi. 2007. Recent changes in the function and frequency of standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics* 11: 437–474.
Hosmer, David and Stanley Lemeshow. 2000. *Applied Logistic Regression*. New York: Wiley.
Hothorn, Torsten, Kurt Hornik and Achim Zeileis. 2006. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics 15*(3): 651–674.
Hundt, Marianne. 2004. The passival and the progressive passive: a case study in layering in the English aspect and voice systems. In Hans Lindquist and Christian Mair, eds*. Corpus Approaches to Grammaticalization in English*. Amsterdam: Benjamins, 79–120.
Hundt, Marianne. 2007. *English Mediopassive Constructions*. Amsterdam: Rodopi.
Hundt, Marianne. 2013. Using web-based data for the study of global English. In Manfred Krug and Julia Schlüter, eds. *Research Methods in Language Variation and Change*. Cambridge: University Press, 158–177.
Hundt, Marianne and Christian Mair. 1999. Agile and uptight genres: The corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4(2): 221–242.
Hundt, Marianne, Lena Zipp and André Huber. 2015. Attitudes towards varieties of English in Fiji: A shift to endonormativity? *World Englishes* 34(3): 688–707.
Hundt, Marianne, Gerold Schneider and Elena Seoane. 2016. The use of the *be*-passive in academic Englishes: Local vs. global usage in an international language. *Corpora* 11(1): 31–63.
Kachru, Yamuna. 2006. *Hindi*. Amsterdam: Benjamins.
Keenan, Edward L. 1985. Passive in the world's languages. In Timothy Shopen, ed. *Language Typology and Syntactic Description*, Volume 1. Cambridge: Cambridge University Press, 243–281.
Kondo, Takako. 2005. Overpassivization in second language acquisition. *International Review of Applied Linguistics in Language Teaching (IRAL)* 43: 129–161.
Kortmann, Bernd and Benedikt Szmrecsanyi. 2009. World Englishes between simplification and complexification. In Thomas Hoffmann and Lucia Siebers, eds. *World Englishes. Problems, Properties and Prospects*. Amsterdam: Benjamins, 263–286.
Leech, Geoffrey, Marianne Hundt, Christian Mair and Nicholas Smith. 2009. *Change in Contemporary English: A Grammatical Study.* Cambridge: Cambridge University

Press.
Lynch, John. 1998. *Pacific Languages: An Introduction*. Honolulu: University of Hawai'I Press.
Maratsos, Michael. 1988. Crosslinguistic analysis, universals, and language acquisition. In Frank E. Kessel, ed. *The Development of Language and Language Researchers. Essays in Honor of Roger Brown.* Hilldale: Lawrence Erlbaum, 121–152.
Matthews, Stephen and Virginia Yip. 1994. *Cantonese: A Comprehensive Grammar*. London and New York: Routledge.
McFarland, Curtis D. 2008. Linguistic diversity and English in the Philippines. In Ma. Lourdes S. Bautista and Kingsley Bolton, eds. *Philippine English: Language and Literary Perspectives*. Hong Kong: Hong Kong University Press, 131–156.
Menard, Scott. 2010. *Logistic Regression: From Introductory to Advanced Concepts and Applications*. Thousand Oakes: SAGE Publications.
Pinheiro, José C. and Douglas M. Bates. 2000. *Mixed-effects models in S and S-PLUS.* New York: Springer.
R Core Team . 2016. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Ransom, Evelyn. 1979. Definiteness and animacy constraints on passive and double-object constructions in English. *Glossa* 13: 215–240.
Rosenbach, Anette. 2007. Animacy and grammatical variation: findings from English genitive variation. *Lingua* 118: 151–171.
Sandahl, Stella. 2000. *A Hindi Reference Grammar*. Leuven: Peeters.
Schachter, Paul and Fe T. Otanes. 1972. *Tagalog Reference Grammar*. Berkeley: University of California Press.
Schneider, Edgar W. 2007. *Postcolonial English: Varieties around the World.* Cambridge: Cambridge University Press.
Schneider, Gerold. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. PhD, University of Zurich.
Schütz, Albert J. 2014. *Fijian Reference Grammar*. Honolulu: Pacific Voices Press.
Seoane, Elena. 2006. Changing styles: on the recent evolution of scientific British and American English. In Christiane Dalton-Puffer, Dieter Kastovsky, Nikolaus Ritt and Herbert Schendl, eds. *Syntax, Style and Grammatical Norms: English from 1500-2000.* Bern: Peter Lang, 191–211.
Seoane, Elena. 2009. Syntactic complexity, discourse status and animacy as determinants of grammatical variation in English". *English Language and Linguistics* 13(3): 365–384.
Seoane, Elena. 2012. Givenness and word order: A study of long passives in Modern and Present-Day English. In Anneli Meurman-Solin, María José López-Couso and Bettelou Los, eds. *Information Structure and Syntactic Change in the History of English*. Oxford: Oxford University Press, 139–163.
Seoane, Elena and Lucía Loureiro-Porto. 2005. On the colloquialization of scientific British and American English. *ESP Across Cultures* 2: 106–118.
Shibatani, Masayoshi. 1988. Voice in Philippine languages. In Masayoshi Shibatani, ed. *Passive and Voice*. Amsterdam: Benjamins, 85–142.
Silverstein, Michael. 1976. Hierarchy of features and ergativity. In Robert M. W. Dixon,

ed. *Grammatical Categories in Australian Languages*. Canberra: Australian Institute of Aboriginal Studies, 112–171.
Strobl, Carolin, Torsten Hothorn & Achim Zeileis. 2009. Party on! A new, conditional variable-important measure for random forests available in the party package. *The R Journal* 1(2): 14–17.
Szmrecsanyi, Benedikt, Jason Grafmiller, Benedikt Heller and Melanie Röthlisberger. 2016. Around the world in three alternations: modelling syntactic variation in varieties of English. *English World-Wide* 37(2): 109–137.
Tagliamonte, Sali & Harald Baayen. 2012. Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24, 135–178.
Thomason, Sarah G. 2013. Innovation and contact: the role of adults (and children). In Daniel Schreier and Marianne Hundt, eds. English as a Contact Language. Cambridge: Cambridge University Press, 283–297.
Wasow, Thomas. 2002. *Postverbal Behavior*. Stanford: CSLI Publications.
Wasow, Thomas and Jennifer Arnold. 2003. Post-verbal constituent ordering in English. In Gunter Rohdenburg and Britta Mondorf, eds. *Determinants of Grammatical Variation in English*. Berlin: de Gruyter, 119–154.
Xiao, Richard, Tony McEnery and Yufang Qian. 2006. Passive constructions in English and Chinese. A corpus-based contrastive study. *Languages in Contrast* 6(1): 109–149.
Zaenen, Annie, Joan Bresnan, M. Catherine O'Connor, Jean Carletta, Andrew Koontz-Garboden, Tom Wasow, Gregory Garretson and Tatiana Nikitina. 2004. Animacy encoding in English: Why and how. In *Proceedings of the ACL-04 Workshop on Discourse Annotation*.
(http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.154.7; accessed 23 November 2017)
Zipp, Lena. 2014. *Educated Fiji English. Lexico-grammar and Variety Status*. Amsterdam: Benjamins.
Zuur, Alain F., Elena Ieno, Neil J. Walker, Anatoly A. Saveliev, and Graham M. Smith. 2009. *Mixed Effects Models and Extensions in Ecology with R.* New York: Springer.