

Noun phrase complexity in Ghanaian English

Thorsten Brato

Department of English and American Studies,
University of Regensburg, Regensburg, Germany

Correspondence

Thorsten Brato, Department of English and
American Studies, University of Regensburg,
Regensburg, Germany.
Email: Thorsten.Brato@ur.de

Abstract

This study compares the complexity of the noun phrase (NP) in Ghanaian English in a real-time perspective. Based on the Historical Corpus English in Ghana (1966–1975) and the Ghanaian component of the International Corpus of English (mainly 2000s), representing the early and late stages of structural nativisation in the dynamic model, NP complexity is measured using five criteria (NP pattern, complexity of the premodifier, length of the pre- and postmodifier and complexity of the postmodifier). The study uses conditional inference trees, taking into account three potential factors: corpus, text type and syntactic function. The results show that over time the complexity of the NP in Ghanaian English has increased, which is interpreted as Ghanaians becoming more proficient and therefore employing more sophisticated structures. Text type is an important predictor; syntactic function plays only a minor role. By adding a diachronic perspective, the study contributes to current models of the evolution of world Englishes.

1 | INTRODUCTION

Several aspects of the noun phrase (NP) have been studied in the context of world Englishes research, such as relative clause formation in Thai (Pingkarawat, 2009) and Ghanaian English (Huber, 2012), the dative alternation in South Asian Englishes (Bernaisch, Gries, & Mukherjee, 2014), genitive alternation in Nigerian English (Akinlotan, 2016) or plural marking in Tswana English (Mohr, 2017). The structural complexity of the NP is at the heart of a number of recent publications. Berlage (2014) presents a large-scale study on spoken and written British and American English and suggests a general framework for studying NP complexity, which will be adopted in a slightly simplified form in the current study, detailed in section 4.2. Brunner (2014, 2017) compares written and spoken data from the British, Singaporean and Kenyan components of the International Corpus of English (ICE) with respect to language contact phenomena and simplification. He shows that there are several effects of the NP structure in the substrate languages (Singaporean English uses more premodified NPs, while Kenyan English uses mainly postmodified NPs – as would be the dominant

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *World Englishes* published by John Wiley & Sons Ltd

patterns in the local language), which he (Brunner, 2017: 247) argues are results of cognitive entrenchment. He further claims that there is a significant effect of register – less formal registers are more likely to be affected by transfer than more formal ones. Regarding simplification, Brunner (2014: 30; Brunner, 2017: 168–169) hypothesises that Singaporean English will use simpler NP structures than British English, and Kenyan English even simpler ones. This, he contends, is due to second language acquisition effects and the fact that Singaporean English has advanced further in the dynamic model of the evolution of postcolonial Englishes (Schneider, 2007) and that speakers and writers in Singapore have a greater choice of complexity levels at their disposal than Kenyans, for which the use of simpler structure is more a functional need (Brunner, 2017: 290).

The study by Schilk and Schaub (2016) compares NP complexity across five components of ICE (Canada, Hong Kong, India, Jamaica and Singapore), also taking into account spoken and written data. They hypothesise that 'NP complexity is mainly influenced by [...] a) syntactic function, b) text type and c) variety' (Schilk & Schaub, 2016: 60). Syntactic function is modelled as a binary factor (subject, other), referring to whether the NP in question functions as a subject or any of the other functions. Complexity is measured on a four-point scale similar to that of de Haan (1993) outlined in section 4.2. They use multinomial regression models and can confirm previous findings that NPs functioning as subjects and in less formal styles tend to be syntactically less complex, while variety – without taking into account interactions – is not a very good predictor. A recent study based on the written sections of ICE Nigeria (Akinlotan & Housen, 2017) researched the impact of grammatical (syntactic) function, length and text type on NP complexity. The authors found that shorter NP constructions, the adverbial function and literary writing are associated most strongly with simpler NPs. Complex NPs are on average longer, function as a direct object and are found in academic writing.

The current study aims to complement this previous work on NP complexity in world Englishes by adding a real-time perspective. Drawing on 4,800 NPs, randomly selected from four genres representing different formality levels, the paper provides a comparison of NP complexity in the early (1966–1975) and late (early 2000s) stages of structural nativisation in Ghanaian English. NP complexity is measured on the basis of four main parameters: (1) the pattern of the NP following de Haan's (1993) classification into unmodified, pre-modified, post-modified and pre- and post-modified NPs; (2) the length of the pre- and post-modification in number of words, as used by Brunner (2014, 2017); (3) the complexity of the premodification (on the basis of three-way distinction into unmodified simple and complex premodifiers); and (4) the structure of the post-modification adapting the classification system outlined by Berlage (2014). Based on the previous research and structure of the NP in the most widely spoken Ghanaian languages (see section 3), we expect the following patterns:

- the more formal the text type, the more complex is the structure of the NP;
- NPs functioning as subjects will be less complex as those in other functions;
- as Ghanaian English has progressed further in the dynamic model, the data from ICE Ghana should show more complex structures than that in HiCE Ghana; and
- complex premodified NPs should be rare, as the common L1s in Ghana only use postmodification.

2 | MORPHOSYNTACTIC VARIATION AND DIACHRONIC DEVELOPMENTS

Until recently, most studies trying to model diachronic morphosyntactic developments to support or refute claims made in the dynamic model had to rely mainly or solely on synchronic data. Most frequently, researchers rely on ICE as their main source of data. One of its key advantages is the comparability across varieties as all components are compiled using identical guidelines and the same corpus design.¹ However, this is also one of its key drawbacks for diachronic studies as it means that researchers have to extrapolate as the data in the corpora stems from the same time period. In other words, it is only possible to compare, for example, a variety that in the dynamic model has advanced to, say, phase 3 to one that has advanced to phase 4 or 5 (such as the studies by Brunner, 2014, 2017 and Schilk & Schaub, 2016) but you cannot compare intra-variety developments so that the results may be contradictory depending on the

type of phenomenon studied. One study (Mukherjee & Gries, 2009), for example, comparing Hong Kong, Indian and Singaporean English finds that there is a correlation between advancement and dissimilarity regarding verb complementation patterns. Another study (Loureiro-Porto, 2016) argues that Indian English should be more different from British English than Hong Kong English (despite both India and Hong Kong being about equally advanced in the dynamic model) in terms of the use of (semi-)modals because it entered the nativisation phase about 100 years earlier.

To overcome these drawbacks and be able to add a real-time perspective, scholars have begun to compile historical and diachronic corpora of Postcolonial Englishes (PCEs) to increase the time depth of possible comparisons across and within varieties of English, applying different corpus designs. The Corpus of Historical Singapore English (Hoffmann, Sand, & Tan, 2012) aims to provide 10-year snapshots from 1951–2011, the Diachronic Corpus of Hong Kong English (Biewer, Bernaisch, Berger, & Heller, 2014) follows the design of Brown and LOB and will include comparable data from 1931, 1961, 1991 and 2011. Another corpus following this design is Phil-Brown, a corpus of historical Philippine English which has been used in relation with ICE Philippines to study diachronic variation in several verb-related phenomena (Borlongan & Dita, 2015; Collins, Borlongan, & Yao, 2014). In the Black South African English context, van Rooy and Piotrowska (2015) use newspaper data from the 1880s to 2010s in their study of real-time change in the use of the progressive. Brato (2018, 2019a) has compiled a Historical of Corpus of English in Ghana (HiCE Ghana; outlined in the following section), which in the present study will be used together with the Ghanaian component of ICE (Huber & Dako, 2013) to study real-time variation and change in NP complexity in Ghanaian English (GhE).

3 | BACKGROUND

Ghana, as of 2016, has a population of 28 million people (Ghana Statistical Service, 2016), who speak about 81 languages (Lewis, Simons, & Fennig, 2013), mostly from the Niger Congo family. Languages in the south mainly belong to the Kwa branch, those in the north are mainly Gur languages. Ewe, spoken in eastern Ghana is a Gbe language. Nine local languages are government-sponsored, which means – among other things – that teaching material is produced in these languages (Huber, 2004: 847). Of the indigenous languages, those subsumed under the heading Akan, belonging to the Kwa branch, are the most widely spoken and serve as a lingua franca across large areas of the country. Other major languages include Ewe, Ga and Dangme (Kwa), spoken mainly in and around the capital Accra, and Dagaare and Dagbani (Gur) in the northern regions. The languages in question share relevant typological features, most notably that head nouns are usually postmodified (for example for numerals or adjectives; Christaller, 1875/1964, pp. 106–110 on Akan; Ameka, 2012, pp. 45–49 on Ewe; and Dryer & Haspelmath, 2013 on the other languages mentioned), a feature shared by most African languages (Heine & Leyew, 2008: 22, referring to work by Heine, 1976). In terms of its development in the dynamic model, Ghanaian English currently falls between the nativisation and endonormative stabilisation phases (Huber, 2014, pp. 87–91). Contact with Europeans goes back at least as far as the late 15th century, but the onset of the foundation phase is 1632, when the British established their first permanent post close to today's city of Cape Coast. Formal colonisation, which marks the beginning of the exonormative stabilisation phase started in 1844 and lasted until independence in 1957. With independence, Ghana moved into the nativisation phase, which is characterised by large-scale linguistic change as more and more of the indigenous population acquire and learn English and transform the language on all linguistic levels, most notably in a distinct accent and a localised vocabulary (Brato, 2019b), but also in preferences for certain morphosyntactic structures (Brato, 2018).

4 | DATA AND METHODOLOGY

4.1 | Data

The present study takes a diachronic perspective on the basis of two corpora. The Historical Corpus of English in Ghana (HiCE Ghana) is a 600,000-word corpus of Ghanaian English from the period 1966–1975. This time frame covers the

TABLE 1 Complexity of the premodifier

No premodification	(DET)		Head	<i>the person</i>
Simple premodification	(DET)	Premod (1 word-class)	Head	<i>the good old days</i>
Complex premodification	(DET)	Premod (2+ word-classes)	Head	<i>the big army vehicles</i>

early years of the nativisation period (Brato, 2019a). Its structure is based on the written-printed categories of ICE (W2A-W2F). It was modified regarding the sizes of the genres. In addition, some genres have finer divisions than those of ICE. These are: press reports, which were subdivided into (a) political news, (b) cultural news, (c) regional news and (d) sports. Similarly, administrative writing was split up into government and non-government subcategories. A further division was made in the 'Creative writing' genre into novels and stories. A category 'Letters to the editor' was added to 'Persuasive writing'. However, this category is not relevant in the current study.

The second corpus is the Ghanaian component of ICE (Huber & Dako, 2013), for which the majority of the data stems from the early 2000s. This period should be considered to representative of the final stages of nativisation as Ghanaian English 'shows the first signs of Endonormative Stabilization' (Huber, 2014: 90). A more detailed discussion of the motivation and theoretical and methodological challenges in the compilation of HiCE Ghana and its comparability to ICE Ghana can be found in Brato (2019a). The two corpora were part-of-speech (POS) tagged in Treetagger (Schmid, 1994), so that potential NP heads could be extracted randomly. As NPs with pronoun heads are usually not (pre-)modified (Biber, Johansson, Leech, Conrad, & Finegan, 1999: 581), these were excluded from the analysis. Similarly, all heads with a proper noun were excluded. In other words, the potential NP heads were collected from those tokens tagged NN (singular or mass common noun) and NNS (plural common noun).

As it was expected that formality of the text genre would play a role in the distribution of the patterns of NP complexity (de Haan, 1993; Schilk & Schaub, 2016) and in order to bring down the number of potential NPs to a manageable size, only NPs from four (sub-)categories representing different levels of formality were taken into account. These are, in descending order of formality: administrative writing, press reports, skills and hobbies, and creative writing. A regular expression was used in AntConc (Anthony, 2018) to extract potential NP heads as well their (potential) premodification. While this regular expression is likely to miss some potential NPs, it was considered versatile enough to provide a good cross-section of the data. All potential NPs together with their context to the left and right were collected. Another set of regular expressions was used first to extract only the context up the start/end of the sentence and then remove anything from the context to the right of the NP head that could not be a postmodification (for example when the head was directly followed by a non-finite or modal verb). Following this first cleaning, the data was laid out in several spreadsheets and randomised and for each corpus and combination the first 700 potential NPs were extracted. These were checked and cleaned manually. Firstly, all tokens falsely identified as potential NP heads by Treetagger were removed. In a second step, the premodification area was checked for correctness. For noun-noun compounds this meant moving the noun identified as NP head to the premodifier column and the noun directly following to the NP head column. Thirdly, the column containing the context to the right of the NP head was analysed to find any postmodifications. In the end, the first 600 annotated NPs from each corpus and genre were analysed, totalling a number of 4,800 NPs.

4.2 | Measurements of NP complexity

Researchers have measured NP complexity in various ways in previous work (Schilk & Schaub, 2016: 62–64). The current study uses five parameters to assess the complexity of the NP in Ghanaian English: (1) de Haan's (1993) classification into simple (unmodified), premodified, postmodified and pre- and postmodified NPs; (2) the complexity of the premodification (not premodified, simple premodification – modification involving a single word class, and complex premodification, in which the premodifier consists of two or more word classes (see Table 1); (3) the length of the premodification in number of words; (4) the length of the postmodification in number of words (Brunner, 2014, 2017);

and (5) the type of the postmodification, using an adapted version of Berlage's (2014, pp. 42–58) classification on the basis of node counts. She proposes that in addition to unmodified NPs there are eight different types in which the NP is modified by a single construction (for example a prepositional phrase) and four types of multiple postmodification (such as involving finite clauses). Initially, 19 different NP types were identified as, a more fine-grained classification for multiple postmodifications was used. In the end, a very broad classification is used in the current study with no further subcategories:

1. non-postmodified;
2. simple postmodification, covering types 2 to 9 in Berlage (2014: 44); and
3. multiple postmodification, covering types 2 to 5 in Berlage (2014: 53) and the additional categories introduced in the annotation of the current data

(1) to (3) exemplify the different levels of postmodification complexity.

- (1) similar steps (HiCE-D2-AdmNGO012) – non-postmodified
- (2) another man, not even the woman's earlier husband (ICE-W2D015) – simple supplement
- (3) another revenge match against AS Aviação on April 11 (ICE-W2C-015) – multiple non-sentential

4.3 | Statistical methods

The study uses conditional inference trees (so-called *ctrees*), using the *R* package *partykit*; (Hothorn & Zeileis, 2015; Hothorn, Hornik, & Zeileis, 2006). Tree-based models have first been introduced to linguistic research by Tagliamonte and Baayen (2012) and have since been used in a range of studies in the world Englishes paradigm (Bernaisch et al., 2014; Koch, Lange, & Leuckert, 2016; Lange & Leuckert, forthcoming). They are an alternative to multiple regression models and particularly useful when the variants are unequally distributed or when the sample size is small, but the number of predictors is large. Furthermore, they are non-parametric and usually avoid overfitting the data. A *cree* shows a series of binary splits, decreasing in their importance until we reach the bar chart or – if the data is metric – the boxplot at the bottom, which shows the distribution of the variables for each final split. Factors, which do not contribute significantly to the model are dropped automatically. Three factors are used here for predicting the complexity of the NP in Ghanaian English:

1. Corpus (HiCE vs. ICE) – a binary variable modelling advancement in time and in the dynamic model. It is expected that NPs in ICE are overall more complex.
2. Category (creative writing, skills and hobbies, press reportage, administrative writing) – an ordinal variable modelling the formality of the text type in ascending order (following Fuchs & Gut, 2015: 378). As text type has been shown to a powerful predictor, we assume that this will also reflect in the current data with NPs increasing in complexity as the formality increases.
3. Syntactic function (Subject vs. non-subject) – a binary variable comparing NPs functioning as subjects to those which do not. We expect that the former are less complex.

To increase the reliability of the *cree*, it is recommended (Levshina, 2015: 297–299) to also run a random forest. A random forest fits a series of *ctrees* (100 in this study), in which the elements are permuted. The resulting variable importance scores indicate how important each of the factors contributes to the overall model. Irrelevant factors will usually get values close to 0.

5 | RESULTS

The following sections all follow the same format. After outlining the descriptive findings for each of the five measures of NP complexity, I will discuss the results of the ctree and random forest.

5.1 | NP pattern

Figure 1 shows the distribution of NP patterns following de Haan's (1993) classification separated by corpus, category and syntactic function.² Unmodified NPs are the largest group of realisations overall and almost equally distributed across the corpora, accounting for 34.7 per cent in HiCE and 36.2 per cent in ICE. There has been a decrease in the premodified variants over time in favour of postmodified NPs, which have increased at about the same percentage. We also note that the complex type has slightly decreased. As expected, there is a clear correlation between category and NP complexity. NPs in the creative writing section are almost twice as likely to be unmodified than in administrative writing, with the other two categories falling in between. While there is almost no difference between the text types with regard to premodification, we note that as formality increases so does the complexity of the NP. For syntactic function, there is a binary split in the data. Unmodified and premodified constructions are more likely found in subject position, while in the other positions more complex NPs are typical. As can be seen from the ctree (Figure 2), the most important factor in the pattern distribution is category. It separates the creative writing and skills and hobbies data on the one hand and press reportage and administrative writing on the other. The second split in the left branch is also by text type. Writing in the skills and hobbies section has neither changed over time, nor is it influenced by syntactic function. Creative writing, which overall shows the least complex constructions, also differs significantly between subjects and non-subjects. On the right hand branch, administrative writing splits of first as it is overall most complex and no further subdivisions may be found. Press reportage has increased in complexity over time. The results from the random forest confirm that category (0.034) is the most important predictor, followed by corpus (0.007) and syntactic function (0.006). However, given how small these values are, it seems that while variation by text type and some change over time may be observed, other factors, not covered here, could be better predictors for the patterns of NP complexity in Ghanaian English.

5.2 | Complexity of the premodification

The second variable in modelling NPs in Ghanaian English is the complexity of the premodification, for which a three-way distinction is used, that into unmodified, simple and complex forms as outlined above. The upper section of Figure 3 confirms our findings from the previous section that overall premodification has decreased over time. Whereas in HiCE about 59 per cent of all NPs are not premodified, this figure has risen to over 65 per cent in ICE. But even in cases in which premodifications are used, these are likely to occur in the earlier data. Given that the dominant Ghanaian languages do not premodify NPs, it may therefore be that writers in HiCE have used more fixed constructions which they have learned in school, whereas writers decades later are more competent language users and may therefore not have to rely on these (see section 6). There is only minimal variation according to syntactic function.

Formality of the text type seems a less good predictor for premodifier complexity. While there are some differences between categories, pointing to the general pattern that there is a split mainly between creative writing and skills and hobbies and the other two, these are less pronounced here. A similar pattern is found for syntactic function, which only varies on a rather small scale.

Figure 4 shows the ctree for this variable. Unlike what was expected from the descriptive findings, the most important split is that between the two less and two more formal contexts. On both sides the next split is in the time dimension. There is no significant difference between creative writing and skills & hobbies for the ICE data, as the latter shows almost the same very large figure (about 69%) of unmodified NPs as HiCE. In the other categories, we find the expected pattern whereby the more formal style also has more complex premodifiers. Although in absolute numbers simple

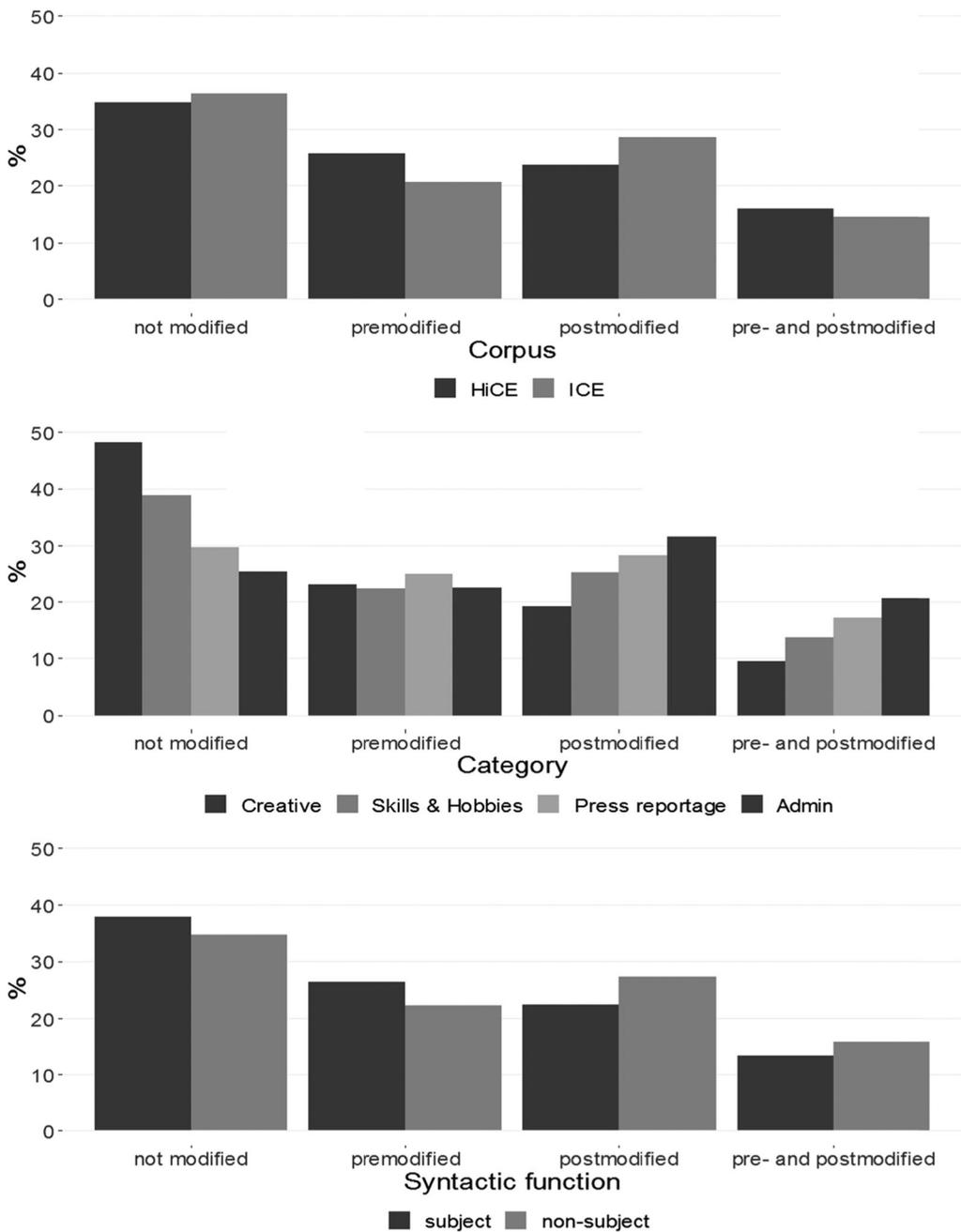


FIGURE 1 Descriptive statistics for NP patterns (de Haan, 1993) by corpus (top), category (centre) and syntactic function (bottom)

premodifications are more likely in administrative writing (38.0%) than press reportage (33.7%) in HiCE, the ctree does not consider this difference significant, whereas in the ICE data we find that in press reportage more premodifications may be found overall and that these are also more complex than in administrative writing. This may be due to the way that headlines are constructed and also the general pattern that administrative writing relies much more strongly on complex constructions, which may be better expressed using a postmodification. As before, variable importance scores

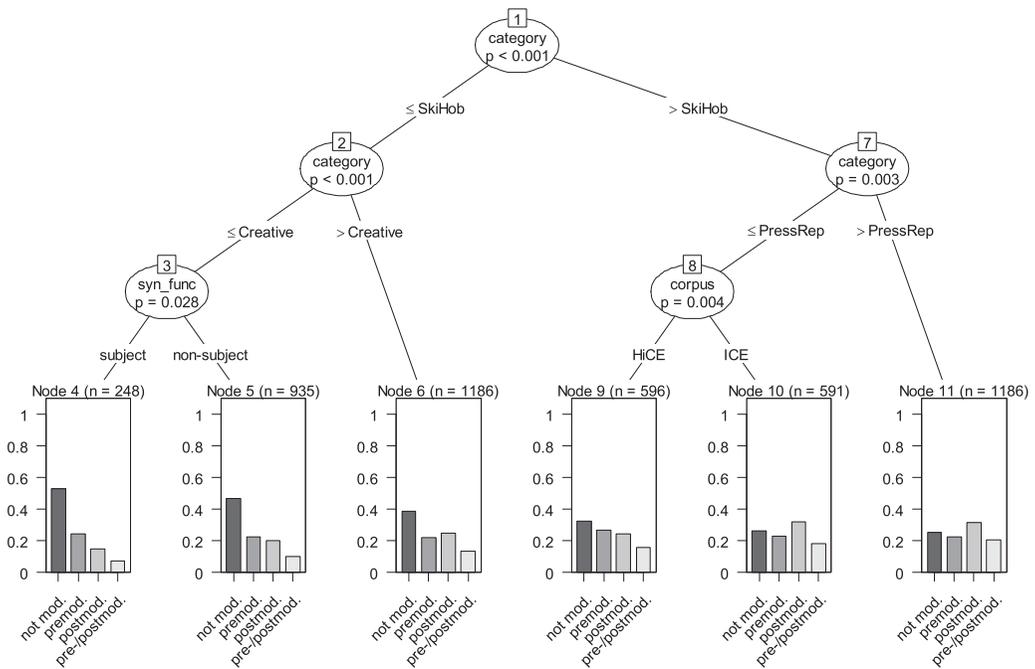


FIGURE 2 Conditional inference tree modelling the dependent variable NP pattern

are extremely low, so the model should be interpreted carefully. Category (0.010) contributes most strongly, followed by corpus (0.006) and syntactic function (0.002).

5.3 | Length of the pre- and postmodification

The following two parameters of NP complexity are metric, namely the length of the pre- and postmodification in number of words. Table 2 shows the mean length and standard deviation (sd) of the premodification. As most of the NPs are not premodified, the numbers are overall small and characterised by their large standard deviations. The patterns from the previous section are confirmed in this data. As there are less premodifications overall in ICE, the average length of the premodification is also shorter. There is some variation with regard to syntactic function along formality in both corpora for NPs not functioning as the subject and for ICE this also holds for the subject function with the caveat that apparently in the admin category all NPs are part of non-subject constructions. As can be seen from Figure 5, the length of the premodification does not yield very fine-grained results. We see once more a split in the middle of the formality levels, which another split in the less formal data, separating the two corpora, although the actual differences are minimal. Unlike expected from the descriptive data, syntactic function does not play a role whatsoever. Also, in terms of variable importance, the model is rather poor (category = 0.015, corpus = 0.006).

Table 3 shows that the length of the postmodification is a much more powerful predictor. First of all, we note an overall considerable difference between HiCE and ICE. The mean value has gone up from 2.2 words in the former to just over three words in the latter. Whereas in HiCE the data is split by category into creative writing, which shows the smallest number of words, skills and hobbies, and Press reportage covering the middle ground, and administrative writing, in ICE the split is in the middle and the differences are overall weaker within the groups. The function in which the NP occurs varies by style, but we also find a diachronic pattern. While there is almost no difference in subject position (possibly due to the lack of data in the Administrative writing section in ICE), for non-subjects the postmodification is almost one word longer in the newer data.

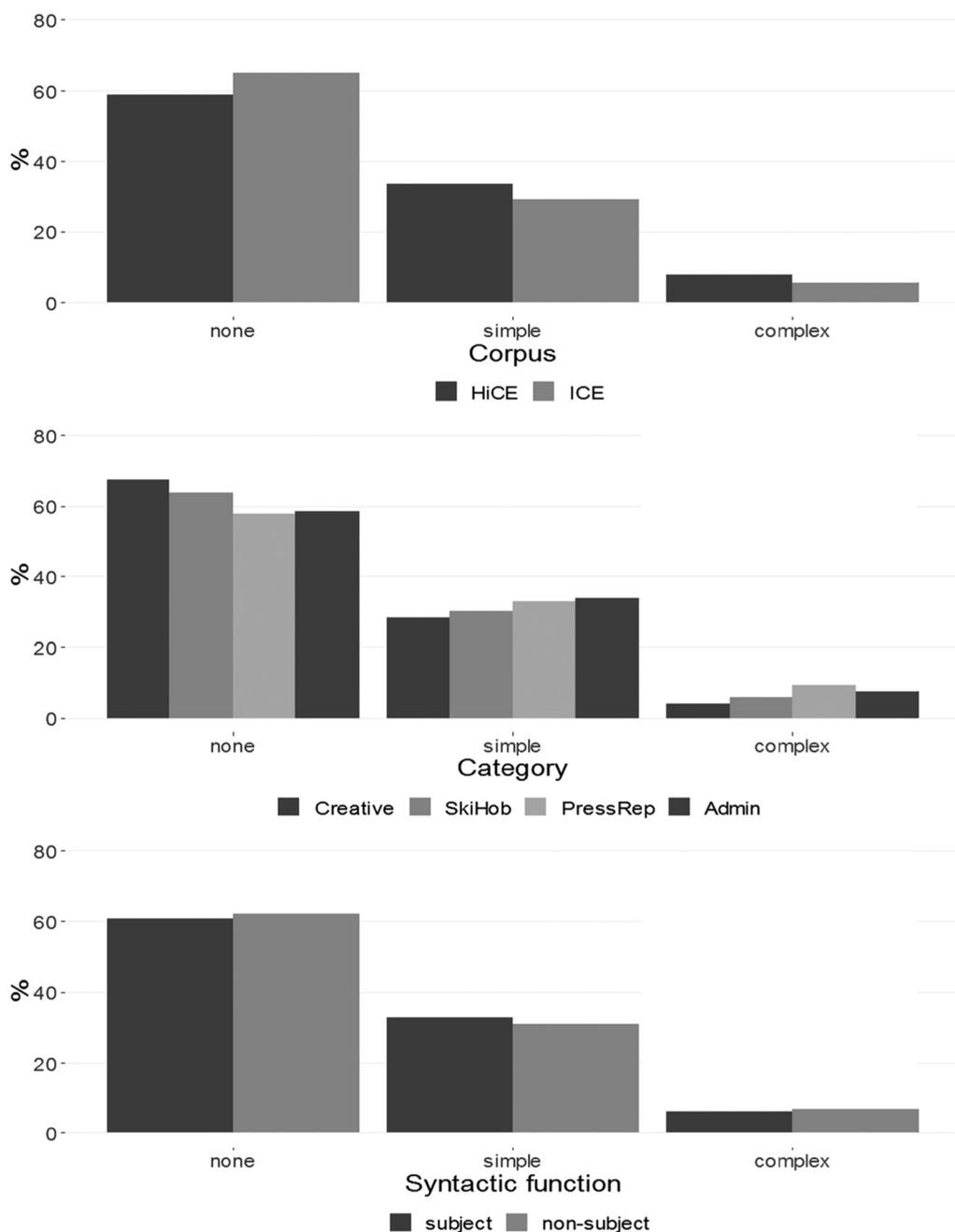


FIGURE 3 Descriptive statistics for complexity of the premodification by corpus (top), category (centre) and syntactic function (bottom)

As Figure 6 shows, the syntactic function in which the NP occurs cannot contribute significantly to the ctree model. Instead, once more the most important split is by formality level, splitting the data in half. Within the less formal data, there is no diachronic difference with regard to skills and hobbies, but corpus contributes significantly to the variation in creative writing, in which the length has increased over time by about 0.5 words, reaching about 1.8 words in ICE. On the right-hand branch, there is – as could be expected from the descriptive data – a split by corpus. While there is no

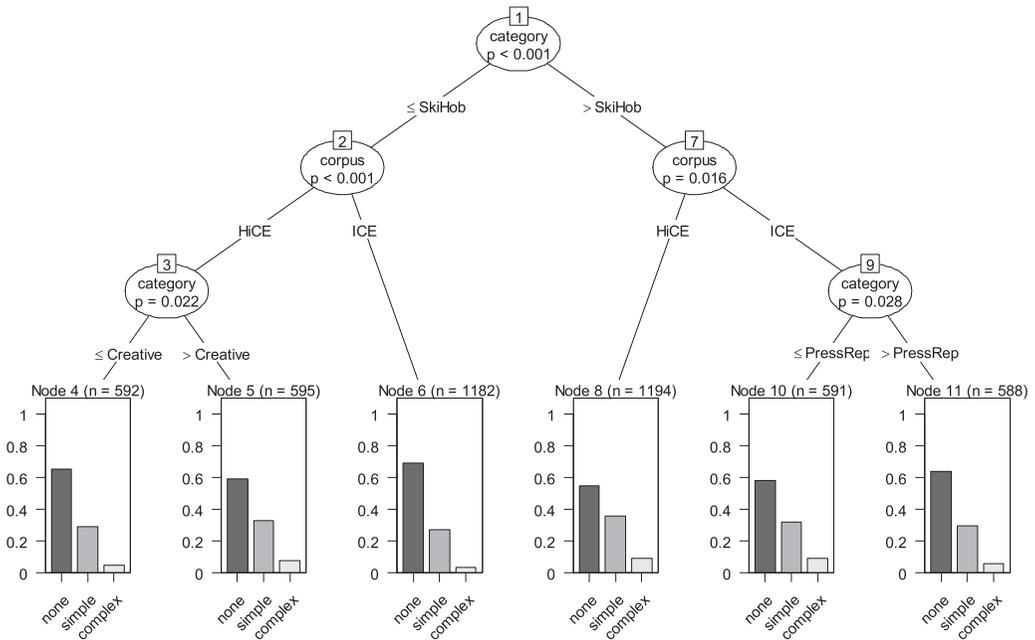


FIGURE 4 Conditional inference tree modelling the dependent variable premodification complexity

TABLE 2 Mean length (sd) of premodification separated by corpus, syntactic function and category

	HiCE		ICE	
	Subject	Non-subject	Subject	Non-subject
Creative	0.46 (0.77)	0.48 (0.78)	0.42 (0.66)	0.39 (0.67)
Skills and hobbies	0.66 (0.88)	0.52 (0.8)	0.36 (0.63)	0.40 (0.66)
Press reportage	0.56 (0.83)	0.64 (1.00)	0.79 (1.28)	0.62 (0.94)
Admin	0.68 (0.83)	0.64 (0.88)	–	0.51 (0.85)

difference in postmodification length between press reportage and administrative writing in ICE, there is a sub-split in HiCE, in which postmodification of press reportage has a length of 2.4 words, compared to about 3.1 in administrative writing. Both predictors contribute highly in explaining the model. The results from the random forest indicate a value of 1.716 for category and 0.587 for corpus.

5.4 | Complexity of the postmodification

The final measure of NP complexity focuses on the postmodification, using a simplified version of Berlage's (2014) classification system, summarised in Figure 7. It shows a decrease in unmodified forms over time of about four percentage points from 59.7 in HiCE to 55.8 in ICE. Differences between the corpora with regard to simple constructions are minimal, but we note that they are slightly more common in HiCE. The data for multiple modification is a counter-image of the first. Here, we find a rather strong increase between the two corpora. Whereas in HiCE only about 16.8 per cent of all NPs were postmodified with multiple constructions, this figure stands at 22.2 per cent in ICE. Once more, the level of formality shows clear-cut differences as to NP complexity. While over 70 per cent of all NPs in the creative writing section are unmodified, this figure decreases continually and is lowest in administrative writing at just under 48 per cent. Simple postmodifications do not differ greatly between the text types, but once more creative writing is

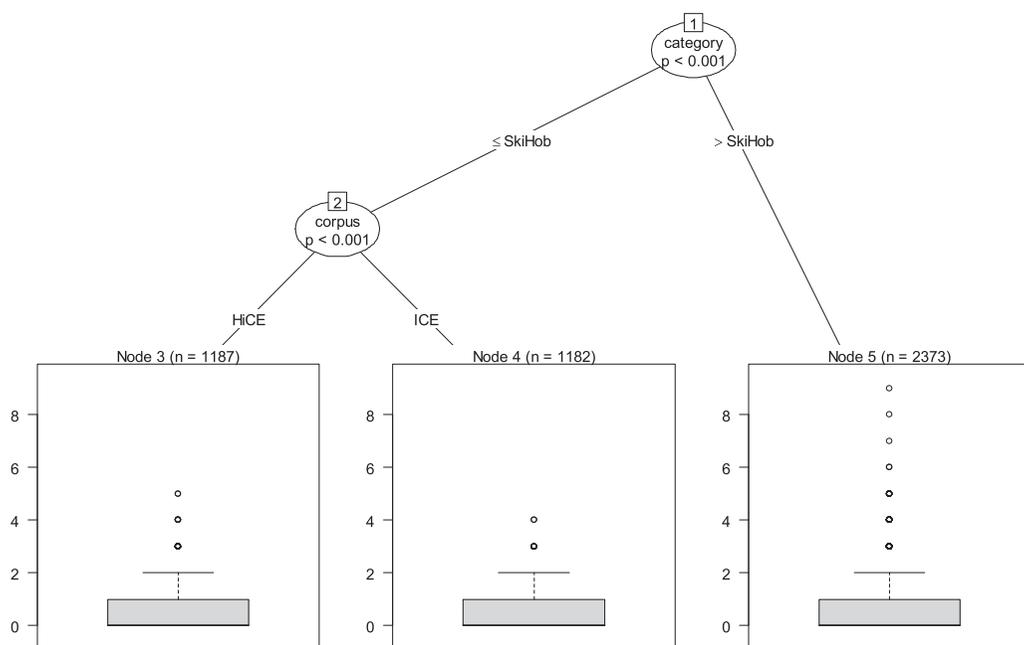


FIGURE 5 Conditional inference tree modelling the dependent variable premodification length

TABLE 3 Mean length (sd) of postmodification separated by corpus, syntactic function and category

	HiCE		ICE	
	Subject	Non-subject	Subject	Non-subject
Creative	0.9 (2.45)	1.32 (2.54)	1.86 (5.13)	1.73 (3.67)
Skills and hobbies	2.07 (5.19)	2.03 (3.52)	1.33 (2.88)	2.12 (3.96)
Press reportage	2.4 (4.85)	2.39 (4.15)	2.94 (5.13)	4.59 (7.11)
Admin	2.63 (4.95)	3.29 (5.05)	–	4.19 (6.50)

the odd-one out with the overall lowest values. Unsurprisingly, the data for multiple postmodifications is the counter-image of the first category. There is a steady increase from under ten per cent in creative writing to almost 28 per cent in administrative texts.

Figure 8 shows the by now familiar split along the middle of the formality of the text types at the top of the ctree. As mentioned above, creative writing is characterised by its large number of unmodified NPs, so that it is separated from skills and hobbies in split 2. Within the latter there is a significant difference between subjects and non-subjects, the former being considerably less complex. The second split in the other branch shows that corpus plays a significant role but that there are no major differences between press reportage and administrative writing, both characterised by the large number of multiply postmodified tokens. This is unlike the older data, in which administrative writing differs greatly by syntactic function in a way that may be expected from previous research, namely that non-subjects are much more complex. The variable importance values gained from running the random forest are small, reaching 0.036 for category, 0.006 for corpus and 0.005 for syntactic function.

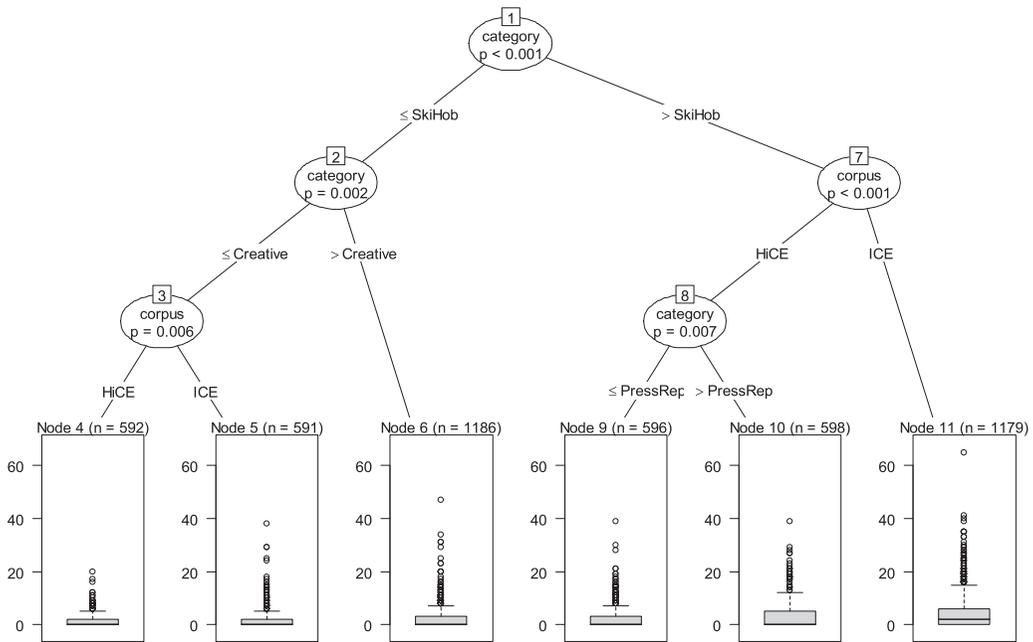


FIGURE 6 Conditional inference tree modelling the dependent variable postmodification length

6 | DISCUSSION

The findings show that NP complexity in Ghanaian English can be measured in different ways and that using the same set of predictors there is considerable agreement in the resulting statistical models based on conditional inference trees and random forests. As was expected based on previous research, the most powerful predictor by far is category, used here to model the formality of the text type. For all five measures of complexity, the first split in the tree divides creative writing and skills and hobbies from press reportage and administrative writing. For NP pattern, it is the only significant factor even on the second level and for both measures of postmodification complexity contributes significantly to further divide the two less formal text types. In total, 12 splits were observed based on this factor. The current study can also confirm that NPs functioning as subjects are less complex than those in other contexts. Overall, however, its importance is comparatively small as it only contributes significantly once the tree was split according to text type and/or corpus. So, in this respect, the current findings run counter to what was reported for Nigerian English by Akinlotan and Housen (2017), who suggest that syntactic function may better model complexity than genre.

The third assumption, namely that the data from ICE Ghana is more complex and that from HiCE Ghana can largely be confirmed. Simpler NP patterns (unmodified and premodified only) account for almost 61 per cent of all tokens in HiCE Ghana, but less than 57 per cent in the ICE Ghana data. There is less premodification in the more recent data, and when it occurs it is marginally shorter, which may be interpreted on the one hand as being less complex, but this should be interpreted as the variety becoming more complex. This is because both parameters of postmodification complexity have increased. The length of the postmodification has gone up from 2.2 words to over three words, the number of NPs that are not postmodified has decreased by four percentage points and that of multiply postmodified NPs now stands at over 22 per cent. All this points in the direction of complexification. The final assumption about the complexity of the premodifier may also be confirmed. The number of complex premodifiers is very low in HiCE Ghana, accounting for 7.7 per cent and has even decreased to only 5.7 per cent in the more recent data.

It makes sense to analyse the reasons for these findings from the last two assumptions together against the background of sociodemographic developments and language policies in the Gold Coast (the colony's name prior

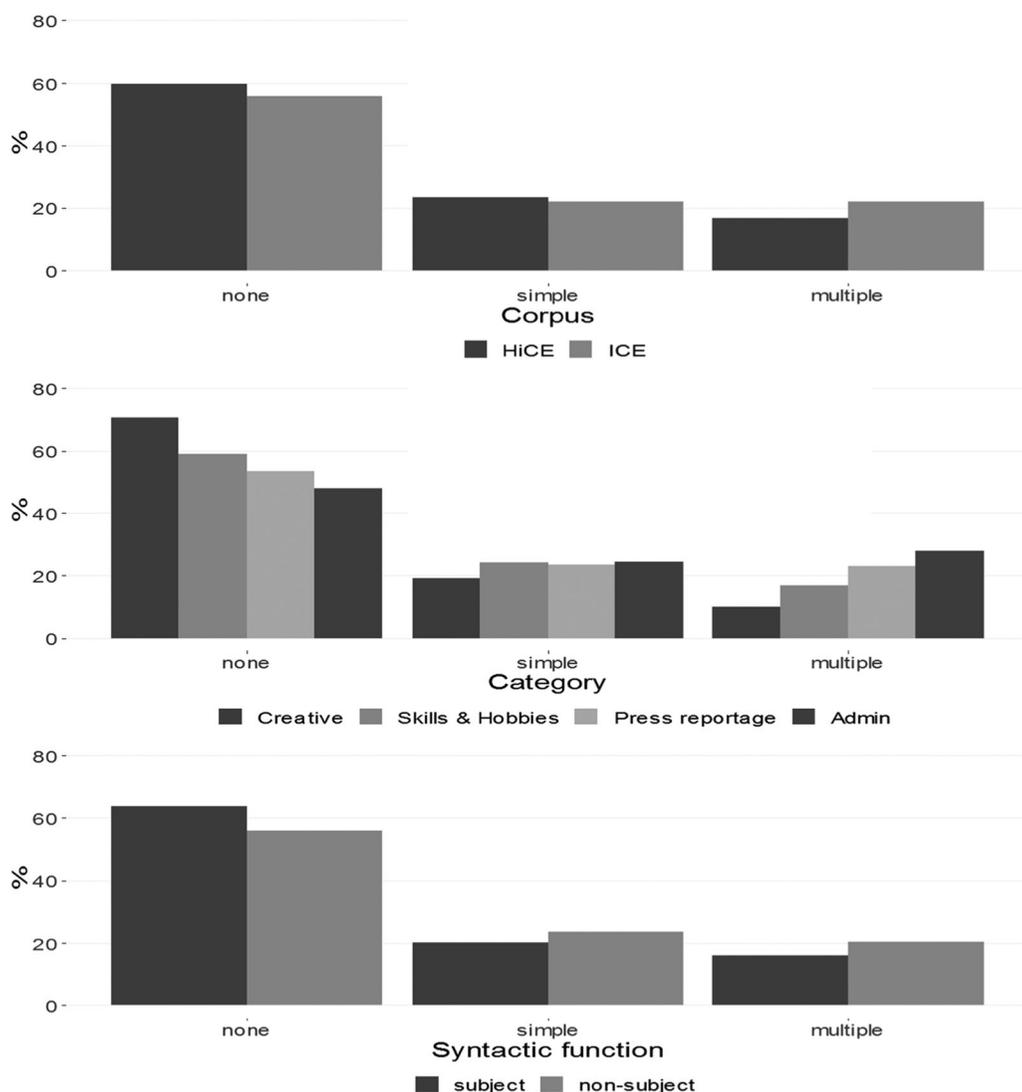


FIGURE 7 Descriptive statistics for complexity of the postmodification by corpus (top), category (centre) and syntactic function (bottom)

to independence) and Ghana during the 20th century. For centuries, the number of schools and, therefore, formally educated Gold Coasters was very small as the British saw little need for educated Gold Coasters. Only a handful would get a job for which English was required, such as working as a clerk or civil servant in the colonial administration. Since the 1880s, there was a constant increase in schools and pupils, not least due to missionary efforts (Jedwab, Meier zu Selhausen, & Moradi, 2018). While in 1920 some 42,000 pupils were enrolled in primary school, the figure for secondary school enrolment stood at a meagre 165. Both figures were to rise substantially and in 1951 more than 300,000 pupils attended primary school, some 6,900 were educated at secondary level. 208 students were enrolled at the University College of the Gold Coast (founded in 1948). Towards the end of the exonormative stabilisation phase and the early nativisation phase (late 1940s to early 1970s) several developments fall together. There is the nationalist movement, campaigning for greater autonomy from the British, which culminated in independence, there is a massive

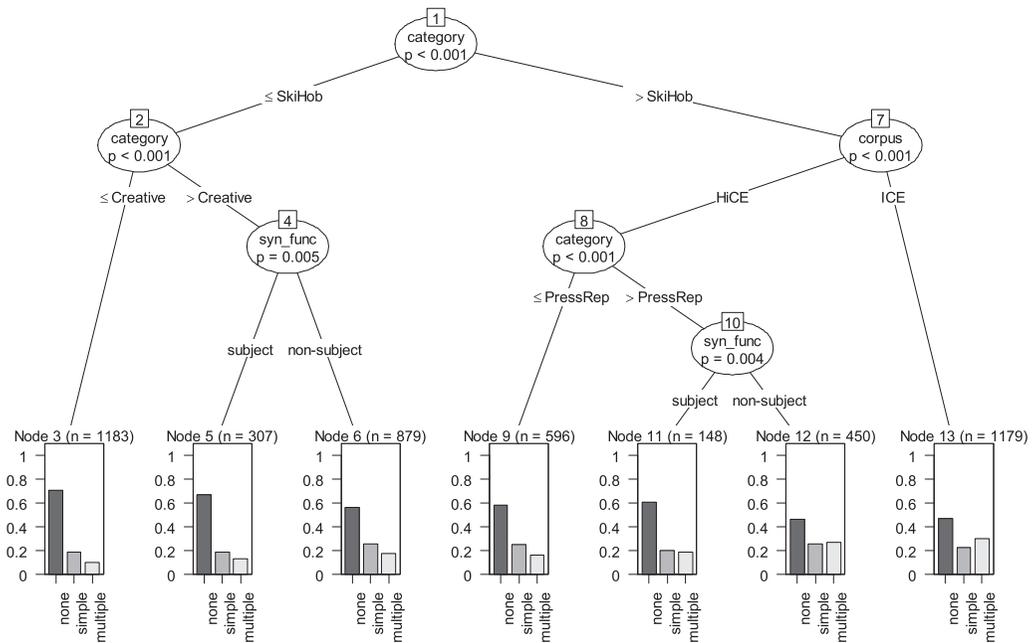


FIGURE 8 Conditional inference tree modelling the dependent variable postmodification complexity

and rapid increase in the population (more than doubling between 1948 and 1970) and there is a large-scale expansion in public primary and secondary education.

In 1966, the onset of the collection period of HiCE Ghana, the population stood at about 8 million, by 1975 this figure had risen to about 9.8 million (The World Bank Group, 2017). Of an adult population of about 4.3 million in 1966, about 20 per cent were educated, which was defined by Sey (1973: 1–2) as having at least gone through ten years of formal education, with only a small proportion moving on to technical or commercial, and even less to tertiary education. He also claims that every educated Ghanaian would speak some English, ranging from a basic form to one that is almost like standardised (British) English. Considering that all writers that became part of HiCE Ghana were at least 18 years old, the education statistics from 1948 (The Government of the Gold Coast, 1950: 18) may be even more relevant. At this point, only 5 per cent of the population aged 10 and above (about 99,000) had attended school until at least the age of ten. This decreases the number of potential writers even further.³ By 1960 about 40 per cent of children of school-going age attended lessons, rising to 72 per cent by 1988 and slightly dropping to 68 per cent in 1998 (Mereku, 2000: 4). In 2016, the literacy rate of the population aged 15 and above stands at 71.5 per cent and (85.7% in the 15–24 range), while over 93 per cent of the primary school children move on to secondary school (UNESCO Institute for Statistics, 2018). We may safely assume that the English spoken and written at the time that is represented by HiCE Ghana was still very strongly influenced by the British model both regarding pronunciation (Huber, 2017) and grammar. Writers included in the corpus almost exclusively fall into stages two and three of what Sey (1973: 14–18) has termed the '[c]orrelation between Standard of Education and the Degree of Bilingualism among E.G.E. [Educated Ghanaian English] speakers'. Ghanaians at Stage 2 have gone through secondary or pre-university training and work as senior clerks, junior civil servants, elementary school teachers or newspaper reporters. Linguistically,

[p]upils, especially at the post-School Certificate stage, are able to write quite sophisticated essays in English, and proficiency in 'registers' associated with special subjects studied for examinations could be very high. [...] Stage Two is the stylistically most ambitious phase in the development of proficiency in the use of English. (Sey, 1973: 15–16)

Those at Stage 3 have a university degree and work as senior officers, lawyers, doctors, secondary school teachers, newspaper editors and other professions that usually require a university degree. The command of English in 'several "registers" [...] is almost native' (Sey, 1973: 16).

As the orientation at the time of the collection period of HiCE Ghana was still strongly exonormative and the number of learners and advanced users of English was so low, written English at the time would be very formal – bookish even – and therefore more complex than the English used today. This is particularly true, when taking into account that following the 1951 Accelerated Development Plan for Education, which saw a rapid and massive increase in primary school and – about eight years later – in secondary school attendance, Ghana was confronted with a lack of qualified teachers throughout the 1950s and 1960s. This in turn led to a stronger vernacularisation of English as the control mechanisms, that is attending one of the prestigious schools in the country to move on socially, no longer worked. However, instead of becoming simpler, NPs in Ghanaian English have become more complex over time. This is not an effect of language contact, despite all major L1s only allowing postmodification of the NP. Rather, we should interpret this a result of increasing proficiency levels following the educational efforts of the last decades, possibly paired with developments in the media (broadcasts by radio, terrestrial and – at a later stage – satellite television) and communication sectors (the abundance of social media). This goes hand in hand with the status that English enjoys, as it seen as a motor of social advancement (Schneider, 2014: 28) or as a medium of interethnic communication (Buschfeld & Kautzsch, 2017: 109). As English has gained ground and the number of proficient speakers has increased so rapidly, the variety became more diverse and its writers more confident in using more complex linguistic structures.

7 | CONCLUSION

The aim of the present study was to add to our understanding of NP complexity in world Englishes by using real-time data. Furthermore, by combining several measures of complexity it tested if there was agreement between these factors and how large it was. It was shown that the general patterns of variation have not changed strongly over time and that there is a very considerable impact of text type, and thus level of formality on NP complexity. The study has also shown that an exonormative orientation and the fact that English was used only by the highly-educated does not mean that the structure of the NP is more complex in HiCE, quite the opposite. I argued that it was rather the larger number of competent and confident users and the more widespread usage triggered the complexification. Whether this holds true for other NP-related features (such as genitive and dative alternation) or with regard to the verb phrase (for example regarding the complementation of verbs or the progressive aspect) is one area of future research. More detailed analyses of real-time change in NP complexity in Ghanaian English should include a larger set of text types, as this factor proved to be the most relevant. Furthermore, the current study only used a binary distinction for modelling the impact of syntactic function. Akinlotan and Housen (2017) show that in the Nigerian context a finer level of granularity is warranted. Thirdly, a much simplified measure of the complexity of the postmodification of the NP was employed here as the number of tokens in some of the categories was extremely small. Increasing the number of NPs and using a more fine-grained level of analysis is certainly necessary. Finally, I suggest that we collect more historical data from other world Englishes so that we can overcome the issues posited by comparing synchronic data from varieties that have developed to different stages in the dynamic model.

ACKNOWLEDGEMENTS

I am grateful to Thomas Kastl for his support in cleaning and categorising the data.

NOTES

¹ Issues of the comparability of ICE corpora have for example been discussed by Mukherjee, Schilk, and Bernaisch (2010).

² A fifth category – other – had to be included for those NP heads which themselves were part of a free wh-clause or gerundial construction. While overall marginal, accounting for only 58 tokens, we note that they have doubled over time. These tokens are no longer taken into account in the following discussion, so from now on the data basis will be 4,742 NPs (2,381 in HiCE, 2,361 in ICE).

³ Very little biographical data could be retrieved for most writers included in the HiCE corpus. Until the foundation of the University College of the Gold Coast in 1948 tertiary education always entailed going abroad. Also, most MA and PhD holders included in HiCE Ghana were educated at least partly in the UK or US. We may assume that the vast majority of texts are by writers who have not been educated abroad.

REFERENCES

- Akinlotan, M. (2016). Genitive alternation in New Englishes: The case of Nigerian English. *Token: A Journal of English Linguistics*, 5, 59–73.
- Akinlotan, M., & Housen, A. (2017). Noun phrase complexity in Nigerian English. *English Today*, 33, 31–38.
- Ameka, F. K. (2012). *Ewe: Its grammatical constructions and illocutionary devices*. Munich: Lincom Europa.
- Anthony, L. (2018). *AntConc (Version 3.5.2)*. [Computer software]. Tokyo, Japan: Waseda University. Retrieved from <https://www.laurenceanthony.net/software>
- Berlage, E. (2014). *Noun phrase complexity in English*. Cambridge: Cambridge University Press.
- Bernaish, T., Gries, S. T., & Mukherjee, J. (2014). The dative alternation in South Asian English(es): Modelling predictors and predicting prototypes. *English World-Wide*, 35, 7–31.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.
- Biewer, C., Bernaish, T., Berger, M., & Heller, B. (2014, April). Compiling The Diachronic Corpus of Hong Kong English (DC-HKE): Motivation, progress and challenges. Poster presented at ICAME 35, Nottingham.
- Borlongan, A. M., & Dita, S. N. (2015). Taking a look at expanded predicates in Philippine English across time. *Asian Englishes*, 17, 240–247.
- Brato, T. (2018). 'Outdooring' the Historical Corpus of English in Ghana. *English Today*, 34, 25–34.
- Brato, T. (2019a). The Historical Corpus of English in Ghana (HiCE Ghana): Motivation, compilation, opportunities. In A. Esimaje, U. Gut, & B. E. Antia (Eds.), *Corpus linguistics and African Englishes* (pp. 120–141). Amsterdam: John Benjamins.
- Brato, T. (2019b). Lexical expansion in Ghanaian English from a diachronic perspective: A structural and semantic analysis. In A. Esimaje, U. Gut, & B. E. Antia (Eds.), *Corpus linguistics and African Englishes* (pp. 260–291). Amsterdam: John Benjamins.
- Brunner, T. (2014). Structural nativization, typology and complexity: Noun phrase structures in British, Kenyan and Singaporean English. *English Language and Linguistics*, 18, 23–48.
- Brunner, T. (2017). *Simplicity and typological effects in the emergence of new Englishes: The noun phrase in Singaporean and Kenyan English*. Berlin: Mouton De Gruyter.
- Buschfeld, S., & Kautzsch, A. (2017). Towards an integrated approach to postcolonial and non-postcolonial Englishes. *World Englishes*, 36, 104–126.
- Christaller, J. G. (1875)[1964]. *A grammar of the Asante and Fante language called Tshi [Chwee, Twi]: Based on the Akuapem dialect with reference to the other (Akan and Fante) dialect*. Farnborough: Gregg Press.
- Collins, P. C., Borlongan, A. M., & Yao, X. (2014). Modality in Philippine English: A diachronic study. *Journal of English Linguistics*, 42, 68–88.
- De Haan, P. (1993). Noun phrase structure as an indication of text variety. In A. H. Jucker (Ed.), *The noun phrase in English: Its structure and variability* (pp. 85–106). Heidelberg: Winter.
- Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *WALS online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from <https://wals.info/>
- Fuchs, R., & Gut, U. (2015). An apparent time study of the progressive in Nigerian English. In P. C. Collins (Ed.), *Grammatical change in English world-wide* (pp. 373–387). Amsterdam: John Benjamins.
- Ghana Statistical Service. (2016). Population statistics. Retrieved from http://www.statsghana.gov.gh/pop_stats.html
- Heine, B. (1976). *A typology of African languages, based on the order of meaningful elements*. Berlin: Dietrich Reimer.
- Heine, B., & Leyew, Z. (2008). Is Africa a linguistic area? In B. Heine & D. Nurse (Eds.), *A linguistic geography of Africa* (pp. 15–35). Cambridge: Cambridge University Press.
- Hoffmann, S., Sand, A., & Tan, P. K. W. (2012). The corpus of historical Singapore English: A first pilot study on data from the 1950s and 1960s. Paper presented at ICAME 33, Leuven.
- Hothorn, T., & Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, 16, 3905–3909. Retrieved from <http://jmlr.org/papers/v16/hothorn15a.html>
- Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15, 651–674.

- Huber, M. (2004). Ghanaian English: Phonology. In E. W. Schneider, K. Burridge, B. Kortmann, R. Mesthrie, & C. Upton (Eds.), *A handbook of varieties of English: A multimedia reference tool* (pp. 842–865). Berlin: Mouton de Gruyter.
- Huber, M. (2012). Syntactic and variational complexity in British and Ghanaian English: Relative clause formation in the written parts of the International Corpus of English. In B. Kortmann & B. Szmrecsanyi (Eds.), *Linguistic complexity: Second language acquisition, indigenization, contact* (pp. 218–242). Berlin: Mouton de Gruyter.
- Huber, M. (2014). Stylistic and sociolinguistic variation in Schneider's nativization phase: The case of Ghanaian English. In S. Buschfeld, M. Huber, T. Hoffmann, & A. Kautzsch (Eds.), *The evolution of Englishes: The dynamic model and beyond* (pp. 86–106). Amsterdam: John Benjamins.
- Huber, M. (2017). Early recordings from Ghana: A variationist approach to the phonological history of an Outer Circle variety. In R. Hickey (Ed.), *Listening to the past: Audio records of accents of English*. Cambridge: Cambridge University Press.
- Huber, M., & Dako, K. (2013). *International Corpus of English, Ghana component*. Gießen: University of Giessen.
- Jedwab, R., Meier zu Selhausen, F., & Moradi, A. (2018). The economics of missionary expansion: Evidence from Africa and implications for development. *CSAE Working Paper Series*, 07, 1–26. Retrieved from <https://www.csae.ox.ac.uk/materials/papers/csae-wps-2018-07.pdf>
- Koch, C., Lange, C., & Leuckert, S. (2016). 'This hair style called as "duck tail"': The 'intrusive as'-construction in South Asian varieties of English and Learner Englishes. *International Journal of Learner Corpus Research*, 2, 151–176.
- Lange, C., & Leuckert, S. (forthcoming). Tag questions and gender in Indian English. In T. Bernaisch (Ed.), *Gender in world Englishes*. Cambridge: Cambridge University Press.
- Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam: John Benjamins.
- Lewis, P. M., Simons, G. F., & Fennig, C. D. (2013). Ethnologue – languages of the world: Ghana. Retrieved from <http://www.ethnologue.com/country/GH>
- Loureiro-Porto, L. (2016). (Semi-)modals of necessity in Hong Kong and Indian Englishes. In E. Seoane & C. Suárez Gómez (Eds.), *World Englishes: New theoretical and methodological considerations* (pp. 143–172). Amsterdam: John Benjamins.
- Mereku, D. K. (2000). *Demand and supply of basic school teachers in Ghana*. National Teacher Education Forum, Winneba. Retrieved from http://wikieducator.org/images/4/4f/DEMAND_SUPPLY_present.pdf
- Mohr, S. (2017). Plural nouns in Tswana English. *World Englishes*, 36, 705–725.
- Mukherjee, J., & Gries, S. T. (2009). Collostructional nativisation in New Englishes: Verb-construction associations in the International Corpus of English. *English World-Wide*, 30, 27–51.
- Mukherjee, J., Schilk, M., & Bernaisch, T. (2010). Compiling the Sri Lankan component of ICE: Principles, problems, prospects. *ICAME Journal*, 34, 64–77.
- Pingkarawat, N. (2009). Relative clauses in Thai English. *Asian Englishes*, 12, 4–33.
- Schilk, M., & Schaub, S. (2016). Noun phrase complexity across varieties of English. *English World-Wide*, 37, 58–85.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.
- Schneider, E. W. (2007). *Postcolonial English: Varieties around the world*. Cambridge: Cambridge University Press.
- Schneider, E. W. (2014). New reflections on the evolutionary dynamics of world Englishes. *World Englishes*, 33, 9–32.
- Sey, K. A. (1973). *Ghanaian English: An exploratory survey*. London: Macmillan.
- Tagliamonte, S., & Baayen, R. H. (2012). Models, forests and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change*, 24, col, 135–178.
- The Government of the Gold Coast. (1950). *The Gold Coast: Census of population 1948 – report and tables*. London: The Government of Gold Coast.
- The World Bank Group. (2017). Ghana population, 1960–2016. Retrieved from <http://data.worldbank.org/indicator/SP.POP.TOTL?locations=GH>
- UNESCO Institute for Statistics. (2018). Ghana: Education and literacy. Retrieved from <http://uis.unesco.org/country/GH>
- Van Rooy, B., & Piotrowska, C. (2015). The development of an extended time period meaning of the progressive in Black South African English. In P. C. Collins (Ed.), *Grammatical change in English world-wide* (pp. 465–483). Amsterdam: John Benjamins.

How to cite this article: Brato T. Noun phrase complexity in Ghanaian English. *World Englishes*. 2020;39:377–393. <https://doi.org/10.1111/weng.12479>