
Article:
Ozon, G. orcid.org/0000-0003-0888-1933, Ayafor, M., Green, M. et al. (1 more author) (2017) The spoken corpus of Cameroon Pidgin English. World Englishes, 36 (3). pp. 427-447. ISSN 0883-2919

https://doi.org/10.1111/weng.12280

A spoken corpus of Cameroon Pidgin English: pilot study

Gabriel Ozon (University of Sheffield), Miriam Ayafor (University of Yaoundé I), Melanie Green (University of Sussex), Sarah FitzGerald (University of Sussex)

Abstract

We report on the construction of a 240,000-word pilot corpus of spoken Cameroon Pidgin English (CPE), a widely-used yet stigmatised and largely uncodified written pidgin/creole variety. The corpus consists of private and public dialogues and monologues, with mark-up and POS-tagging. Text categories and the proportions of monologue and dialogue are guided by those of the International Corpus of English project, which makes the corpus immediately comparable with existing corpora of post-colonial varieties of English. We discuss the extent to which this corpus can be regarded as an ICE component, and illustrate the relation between CPE and standard Nigerian and Cameroonian varieties of English in Africa by means of case studies employing ICE-NIGERIA (Wunder et al. 2010) and the Corpus of Cameroon English (Tiomajou 1993; Nkemleke and Mbangwana 2007). The main challenge of the compilation stage of the CPE corpus has been the development of a systematic orthography. The project has also necessitated the development of a designated tagset for CPE, which has been adapted from the CLAWS5 tagset. Manual tagging of selected texts has enabled training of the Tree Tagger (Schmid 1994), with automatic tagging tests producing positive results (over 90% accuracy). The two-year project is funded by a British Academy/Leverhulme small grant (ref. SG140663). On completion of the project in summer 2016, the recordings and texts have been deposited with the Oxford Text Archive.

1. Introduction[1]

Cameroon Pidgin English (CPE) is an expanded pidgin/creole spoken in some form by an estimated 50% of Cameroon's 22,000,000 population (Lewis et al. 2014), primarily in the anglophone west regions, but also in urban centres throughout the country. As a primarily spoken language, CPE has no standardised orthography, but enjoys a vigorous oral tradition,

[1] This paper was first presented at the ICAME 36 conference in Trier, 27-31 May, 2015. We thank the members of that audience for their comments, as well as the editors of this special issue, Robert Fuchs, Ulrike Gut and Gerry Nelson. We are also grateful to an anonymous reviewer, whose suggestions helped us to improve the paper. We remain responsible for any errors or omissions.

not least through its presence in the broadcast media. CPE, however, has resisted close documentation due to its stigmatised status in the face of French and English, prestige languages of Cameroon, where it also co-exists with an estimated 280 indigenous languages (Lewis et al. 2014).

The present paper reports on the construction of a 240,000-word pilot corpus of spoken Cameroon Pidgin English (CPE). The corpus consists of private and public dialogues and monologues, with mark-up and POS-tagging, and is aimed at providing a resource for linguistic description and comparison, as well as offering the potential for codification, destigmatisation and the development of literacy materials.

The paper is organised as follows. Section 2 provides an overview of existing research on CPE, with a focus on its sociolinguistic status, sets out the motivations and potential uses of the corpus, and describes the areas of expertise of the project investigators. Section 3 addresses the design of the project, which is aimed at comparability with existing corpora of postcolonial varieties of English, and discusses the challenge of achieving representativeness in such a complex linguistic environment, addressing the extent to which this corpus can be regarded as an ICE component. Section 4 describes the compilation of the corpus (recording, transcription and annotation), with a particular focus on the challenge of developing an accessible and systematic orthography for this largely unwritten variety. Section 5 explores the process of devising a tagset for CPE (adapted from CLAWS5), and initial results of automatic tagging tests using the Tree Tagger (Schmid 1994). Section 6 explores the comparability of the CPE corpus with ICE-NIGERIA (Wunder et al. 2010) and the written Corpus of Cameroon English (Tiomajou 1993; Nkemleke and Mbangwana 2007) by means of two short case studies. Section 7 concludes the paper with a summary of the progress of the project, the key challenges, and future developments.

2. Introducing the Spoken Corpus of CPE

The present section provides background on CPE, with a focus on its sociolinguistic status (§2.1), introduces the project in terms of the motivations and potential uses of the corpus, and describes the areas of expertise of the project investigators (§2.2).

2.1. CPE

CPE occupies an important place in a highly complex linguistic ecology. There are an estimated 280 living languages in Cameroon (Lewis et al. 2014), which is one of the most linguistically complex regions in Africa, at the intersection of three of the major language families of Africa: the Afroasiatic family, the Nilo-Saharan family and the Niger-Congo family. In addition, historical contact with German and Portuguese, and with English and French as official languages further adds to the complexity of the contact setting.

A number of authors have published research on the sociolinguistic situation in Cameroon, including Todd (1982), Koenig, Chia and Povey (1983), de Féral (1989), Wolf (2001), Simo Bobda and Wolf (2003) and Menang (2004). However, Schröder (2003) provides the most comprehensive recent study of the sociolinguistic context of CPE. Schröder's research was based on qualitative data from 66 interviews and quantitative data from approximately 2,000 questionnaire respondents, and was carried out at 13 educational establishments: eight high schools and five universities covering eight of the ten administrative areas of Cameroon. The participants were teachers and students in form 5 and above, and approximately 50% were anglophone and 50% francophone (Schröder 2003: 28-37).

Schröder found the highest concentration of proficient CPE speakers in the Anglophone regions, but established nevertheless that CPE is used in some form by a substantial proportion of speakers across Cameroon. Schröder (2003: 85) also found that the 0-15 age group had the lowest proportion of CPE speakers (50%), while the 50+ age group had the highest proportion (83.3%). In terms of attitudes to CPE, Schröder (2003: 54-58) reports a widespread view that CPE is detrimental to the acquisition of 'good English', a view expressed by both anglophone and francophone CPE speakers in her study. Schröder (2003: 64-70) also reports participants' views that CPE is contributing to the endangerment of Cameroon's indigenous languages, despite some ambivalence to multilingualism.

Schröder distinguishes the varieties of CPE according to a number of parameters: regional (anglophone vs. francophone), urban vs. rural, social varieties and situational varieties. With respect to anglophone and francophone varieties, Schröder (2003: 90-98) reports that CPE speakers can often distinguish the two varieties, and mentions phonological, lexical and morphosyntactic differences reported by her participants. With respect to urban and rural varieties, she comments that urban varieties are more likely to show influence from the official European languages, while rural varieties are more likely to show influence from

indigenous languages (Schröder 2003: 101). Indeed, some of Schröder's participants described the rural CPE as 'unadulterated', recognising that urban CPE is more likely to be influenced by English. Urban varieties thus tend more towards the acrolect, while rural varieties are more basilectal. The urban-rural distinction also overlaps with social variation, particularly in relation to education. Education in rural areas tends to be limited to the primary level, while secondary and tertiary educational establishments are located in urban areas. With respect to situational variation, Schröder (2003: 115) observes that the key factor is accommodation to the interlocutor's language preferences and ability, and that educated CPE speakers may have command of a range of lects and use whichever is most appropriate to the situation.

In terms of the functions of CPE, Schröder explores its use in a range of domains. For example, in the domain of education, where the use of CPE is explicitly prohibited, CPE is rarely used between teachers and their students, but it is widely used among the students as a marker of in-group status. In the domains of mass media and politics, Schröder reports that CPE is absent from the official mass media (radio, TV and newspapers), although it is widely used for unlicensed radio broadcasts (predominantly in the anglophone regions), and while it is not used for printed political materials, it is used in political campaigns. In the domain of administration, Schröder reports that educated Cameroonians rely on the official languages for these purposes, but that less educated Cameroonians might use CPE in this domain, a speculation that receives some support from the documentary film *Sisters in Law* (Ayisi and Longinotto 2005). In the domain of trade, Schröder's findings corroborate Ayafor's (1996: 54) statement that CPE is the most widely spoken language 'in market places all over the country'. However, this is more clearly the case for anglophones (Schröder 2003: 151).

Within the ICE programme, Schröder's findings bear similarity to those of Hackert (2010). In her overview of language use in the Bahamas, Hackert reports that the standard English language is 'subject to encroachments from the creole in a number of domains' (2010: 44). For example, while the language of parliamentary debate and administration remains standard English, political speeches evidence a substantial amount of mixing; and whereas newspapers are still written in standard English, the creole is often employed as a stylistic device.

Schröder concludes her study with a discussion of the pros and cons of a national language for Cameroon, and it is striking to note that 29.1% of her participants identify CPE as the

most suitable candidate, although the majority (34.2%) responded that there was no Cameroonian language with potential national language status (Schröder 2003: 196). Schröder attributes the position of CPE in this survey to its relative ethnic neutrality (although many francophones consider it an anglophones' language) and its geographical spread, but also points out a number of drawbacks, among them the absence of a standardised orthography and, most significantly, the low social status of CPE, which is widely considered a form of 'broken English', not a 'proper' language, and a medium of communication for the uneducated (Schröder 2003: 206-207).

Structural descriptions of CPE date back to the 1960s and early 1970s (e.g. Schneider 1966, Todd 1969, Gilman 1972, Mbassi-Manga 1973). Three pedagogical grammars have also been published since the 1960s (Dwyer 1966, Todd 1991 and Bellama et al. 2006). More recent years have seen the publication of a dictionary with a short section of grammatical notes (Kouega 2008), a short grammatical sketch (Ayafor 2004), a short phonological sketch (Menang 2004), a collection of papers focusing on structural and sociolinguistic issues (Anchimbe 2012), and most recently, Nkengasong's (2016) volume, which provides a discussion of the socio-cultural context of CPE and its orthography, as well as a brief overview of word classes and sentence types and a collection of proverbs. Ngefac (2014) provides a historical overview of CPE, and a number of papers address the issue of orthography (e.g. Ayafor 1996, Sala 2014). A comprehensive descriptive grammar, which draws its data from the current corpus project, will shortly go to press (Ayafor and Green, in press).

2.2. Project

In light of the sociolinguistic context described above, a corpus of spoken CPE is motivated by a number of potential applications: language description/codification, linguistic investigation more generally, comparison, and ultimately destigmatisation.

With respect to description and codification, the pilot corpus allows linguists to identify and describe recurring grammatical patterns, as well as the phonology of the language (given the availability of sound files to be deposited with the text files). While the size of the pilot corpus is not sufficient for lexical studies, it nevertheless allows for the identification of high-frequency lexical phenomena, as we discuss in §6. In terms of codification, the pilot corpus

has also informed the first comprehensive descriptive grammar of the language (Ayafor and Green, in press).

With respect to linguistic investigation more generally, the corpus provides an exceptional resource for the study of general/theoretical linguistics, creolistics, typology, language contact and change, sociolinguistics and discourse analysis. It also allows comparison of CPE with other pidgin/creole languages, other Cameroonian and West African languages, and other varieties of post-colonial English, as illustrated by the case studies below (§6).

In terms of practical applications, the corpus may ultimately offer the potential for developing CPE literacy materials, thus contributing to language planning in the country, particularly in education, where two different exocentric norms are competing, and where CPE is highly stigmatised.

This project relies on the expertise of a linguistically trained native speaker of Cameroon Pidgin English, who also specialises in literacy; an expert in the grammar of African languages; a corpus linguist, and a team of research assistants.

3. Design

The present section sets out the objectives of the project in terms of design principles, as well as challenges.

3.1. Representativeness, comparability, balance and sampling

The ultimate design principle in corpus building is representativeness, with the objective that the findings emerging from the corpus will be generalisable to the larger population of which the corpus represents a sample. The corpus should thus contain representative variation in terms of region, other languages spoken, age, gender, educational background, and so on.

In the Cameroonian context, our objective was to obtain a representative sample along the following dimensions of variation: age, gender, ethnic group/L1(s), level of education, medium of education, and language(s) used at home and at work. This information was collected by means of a participant questionnaire completed before the recording, which allowed us to determine whether the speaker's variety might be expected to lean towards the basilect, mesolect or acrolect. In addition, speakers were asked a set of questions at the end of

their recording sessions concerning their attitudes to CPE. Naturally, this speaker metadata is not available in the case of radio broadcasts.

In their discussion of ICE-Fiji, Biewer et al. (2010:5) describe the challenge of building a corpus which is at once representative of current language use in a particular postcolonial scenario, and at the same time comparable to all other ICE corpora. Furthermore, comparability and representativeness may often pull in different directions, as Leech (2007) points out:

> While it makes sense to achieve success in both representativeness and comparability, there is a sense in which these two goals conflict: an attempt to achieve greater comparability may actually impede representativity and vice versa. (Leech 2007: 142)

We discuss this in more detail below (§3.3). In terms of balance, the CPE pilot corpus was designed according to the same criteria as the spoken component of the International Corpus of English (ICE) project (Nelson 1996). In addition to contributing to the representativeness of the corpus in terms of private and public uses of language, this also ensures direct and immediate comparability with the ICE subcorpora (Table 1).

CPE		**texts**	**words**	**%**
dialogues	private	26	78,000	33%
	public	21	63,000	26%
monologue	unscripted	18	54,000	23%
	scripted	15	45,000	19%
		TOTAL	240,000	

Table 1: Proportions of text categories in CPE pilot corpus

3.2. Design challenges

The intricacy of the language ecology in Cameroon makes identifying criteria for representativeness a challenge: although the project targets CPE 'native speakers', there is considerable variation as a consequence of the complex multilingual environment, in which monolingual speakers are the exception rather than the norm. It follows that identifying a 'native speaker' is not straightforward; for example, someone might use CPE proficiently on a daily basis in certain domains, but may not have spoken CPE as a child and/or may not use

it in the home. Due to these complexities, we relied on the judgement of the research assistants, whose CPE language expertise was sufficient for the identification and selection of proficient speakers.

Given that this corpus was built with the aim of providing a dataset comparable with ICE corpora, ICE guidelines for the selection of speakers were initially considered. These require that the users to be represented in an ICE corpus are 'adults of 18+ who have received formal education through the medium of English to the completion of high (secondary) school… (Greenbaum 1991: 3). Moreover, speakers or writers (…) must be 'natives', i.e., they must have either been born in the country or moved there early in their lives and received their education through the medium of English there (Nelson 1996: 28).

In the complex post-colonial linguistic context described above (§2.1), these criteria for the selection of speakers have a number of consequences:

(i) as already mentioned, it is hard to define what a 'native' speaker of a pidgin/creole variety is (bi-, tri- and multilingualism are widespread in typical contact scenarios);

(ii) the education requirement would considerably reduce the potential number of participants;

(iii) besides the expected inter-speaker variation, intra-speaker variation is also prevalent: even educated users of English deploy a variety of speech forms in accordance with changing situational factors. Accommodation and situational variation are central to post-colonial varieties (§2.1; §5.4).

In view of the above challenges, our project attempted to navigate a course between comparability and representativeness. In line with the approach adopted by other ICE teams, we aimed at capturing CPE as used by competent speakers 'regardless of whether they [were] first or second (or third) language users of English' (Mukherjee et al. 2010). We relied on our field research team (which consisted of CPE native speakers) to make the final decision about whether to include/exclude certain speakers.

3.3. Fit with other ICE corpora

In a number of respects, the CPE corpus can be regarded as an ICE component, given the following formal similarities:

- It has been designed and sampled to be representative (§3.1).
- It has the same proportions of public/private spoken language) (§3.1).
- It has been annotated with the existing ICE mark-up scheme (§4.3; Appendix 2)
- It has been tagged using (a customised version of) CLAWS (§5.1; Appendix 3)

On the other hand, the CPE is distinct from the ICE corpora in the following respects:

- It contains spoken language only, and is thus only comparable with the spoken components of the ICE corpora.
- CPE is predominantly an L2 variety.
- As already mentioned, speakers of a pidgin/creole language typically show greater intraspeaker variation often as a consequence of situational variation.

These and other, similar issues were reported on in *ICAME Journal* volume 34 (2010) dedicated to ICE Age 2. Our CPE corpus is thus more closely comparable to the new generation of ICE corpora, inasmuch as (i) they are all ESL corpora of New Englishes, and (ii) the data were collected post 2000.

4. Compilation

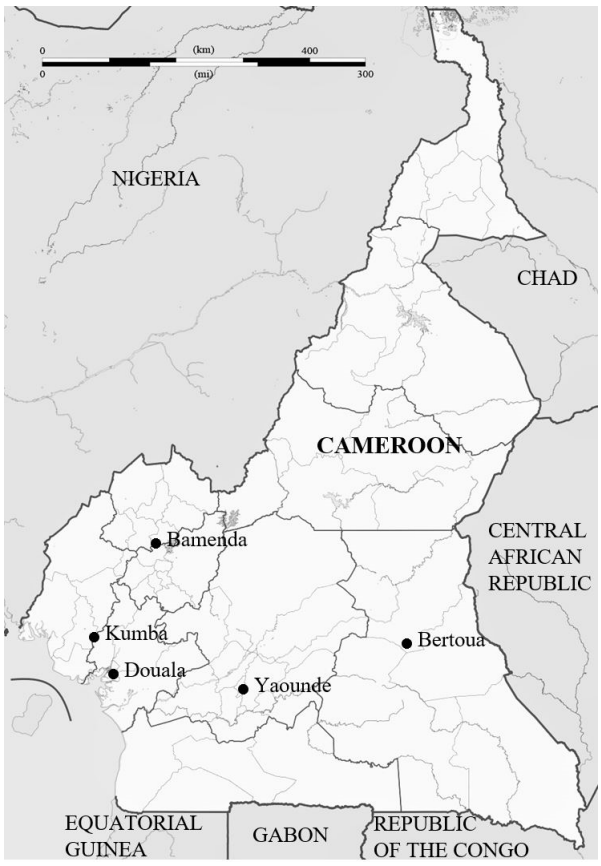
The present section describes the compilation stage of the project, including recording locations, participant recruitment, transcription and annotation.

4.1. Data collection

Data was collected from five regional headquarters of the ten administrative regions that make up Cameroon. The five regional headquarters initially chosen were Bamenda in the North West Region, Bertoua in the East Region, Douala in the Littoral Region, Kumba in the South West Region, and Ngaoundere in the Adamawa Region. Unlike the other cities, Kumba is not the seat of government in the South West Region but was selected for two particular reasons. First, Buea, the regional capital, is very close to Douala and so it was judged that the variety of CPE spoken there may not differ significantly from that spoken in Douala. Secondly, Kumba is a business centre inhabited by many Igbo people from Nigeria, hence the potential influence of Nigerian Pidgin in this town made it a potentially interesting location.[2]

[2] Peter and Wolf (2007:6) attribute the close linguistic similarity (in both pronunciation and grammar) between Nigerian Pidgin English and CPE to the fact that they share a common (linguistic) history of forty years of joint

These five regions are representative of regional variation in Cameroon in the sense that they form the four cardinal points of the country, covering the greater north, the east, the south and the west (Fig 1). Unfortunately, at the time of collecting the data, Cameroon was undergoing serious attacks from the radical Islamic group Boko Haram from neighbouring Nigeria. These attacks targeted the north of the country and travel to Ngaoundere was therefore considered to pose a serious risk to the safety of the researchers. We therefore substituted Yaoundé for this location. However, efforts were made to record members of the Muslim community in Yaoundé, who are Hausa speakers, since they are likely to be representative of the CPE speech that we would have obtained from Ngaoundere where a substantial proportion of our target participants would have been Hausa-speaking Muslims.



administration (see Huber, 1999: 57, 119–29; Holm, 1989: 410–12), and that their geographical proximity allows for cross-border interchange between Pidgin speakers from Nigeria and Cameroon. Furthermore, Peter and Wolf (2007) claim that at the time there were an estimated three million Nigerians living in Cameroon, a number which is sure to be much higher presently.

Figure 1: Map showing recording locations[3]

Participants for the study were recruited either via personal contacts of the researchers or by approaching them directly in the field. Ethical procedures were followed by means of the distribution of standard information forms and the collection of consent forms. Participants were selected according to the sampling criteria outlined above (§3). Metadata on age, gender, educational background and linguistic background was collected by means of a questionnaire that was completed by the researcher prior to the recording.

Data collection for the spoken corpus of CPE was divided into sixteen slots of 30-minute digital recordings in each location. These sixteen recordings consist of (a) five private dialogues including four face-to-face conversations and one phone call, (b) four public dialogues made up of three radio phone-ins or interviews and one conversation taking place in a public location such as a market, restaurant or barber shop, (c) four unscripted monologues consisting of two personal narratives and two demonstrations (e.g. 'How to build a house'), and finally (d) three scripted monologues including one news broadcast, one radio sermon and one live religious sermon or public lecture.

This recording schedule was repeated in each of the five locations. Sixteen 30-minute slots in five locations produced eighty slots containing 2,400 minutes or 40 hours of recorded CPE speech. Sound files were saved in both .wav and .mp3 formats. After completing the recordings in each location, the transcription of those recordings was completed before proceeding to the next location.

4.2. Transcription

Transcription procedure was outlined in a field manual prepared by the investigators and distributed to the research assistants. The field manual emphasises the necessity for accurate transcription, including disfluencies. Because of the Observer's Paradox, each transcription began about three minutes after the start of the recording, and stopped when a target number of slightly over 3,000 words had been reached.

[3] Map by Flappiefh - Own work from: NASA Shuttle Radar Topography Mission (SRTM3 v.2) (public domain); Vectors: DIVA-GIS., CC BY-SA 4.0, http://tinyurl.com/homlymx.

Due to the absence of a standardised orthography for CPE, it was necessary to (a) develop an orthographic system to be included in the field manual, and (b) train the research assistants in using this system. There have been a number of proposals for a CPE orthography, as summarised most recently by Sala (2014). In her various publications, Todd relies on a transcription-based orthography, an approach also advocated by Mbangwana (1983), Ayafor (1996) and Sala (2014). While Ayafor (1996) suggests the use of accents for the representation of different vowel qualities, we did not adopt this proposal for the current project, since accents are conventionally used in linguistics publications to mark tone. The orthography adopted for this project is based on that developed by Ayafor (2014) (Appendix 1). The orthographic system was kept under review during the transcription stage of the project, and a regularly updated spelling guide was produced. Post-checking monitored for intra- and inter-transcriber consistency with respect to the spelling guide provided in the field manual.

4.3. Annotation

The annotation section of the field manual was adapted from ICE guidelines for spoken texts (Nelson 2002), and standard mark-up symbols were used to denote text unit, speaker identification, overlapping speech, unclear words, uncertain transcriptions, anthropophonics, editorial comments, foreign words and indigenous words (Appendix 2).

The first stage of annotation required the segmentation of transcribed texts into utterances/text units, some but not all of which corresponded to speaker turns. Each utterance was given a speaker identification code. Mark-up was added for overlapping speech, unclear words and uncertain transcriptions, anthropophonics (e.g. 'laughter') and editorial comments (e.g. 'break in recording'). Words from European languages (i.e. English or French) were marked as 'foreign' where the transcriber judged that the expression was a loanword for which a near-synonym exists in CPE, or where the speaker was code-switching into English/French. Words from indigenous African languages were marked as 'indigenous' according to the same criteria. Naturally this approach has its limitations, as judgements may be subjective.

Speaker metadata was compiled into a database with a view to including this information in file headers.

4.4. Compilation challenges

The main challenges encountered during the compilation stage of the project were (a) access to participants or data, (b) poor sound quality of certain recordings, and (c) ensuring a consistent orthographic representation of CPE.

Access to participants was difficult in certain locations due to high levels of public anxiety resulting from terrorist activity, which made people nervous about being approached by strangers. In other locations, participation was refused due to the social stigma associated with CPE. In particular, highly educated CPE speakers such as university lecturers would often refuse to participate because they did not want to be recorded speaking CPE. A further challenge was the unavailability of CPE radio broadcasts by radio stations in certain francophone regional locations. This problem was circumvented by substituting nationally-available broadcasts from radio stations in other locations.

In terms of sound quality, it was particularly difficult to eliminate distracting background noise in public recording locations. In addition, the unavailability of digital radio broadcasting in Cameroon entails that broadcasts cannot be recorded directly from the radio, which also results in background noise on some of these sound files.

However, the most significant challenge encountered during the compilation stage was ensuring a consistent orthographic representation of CPE. Given the absence of a standardised orthography, the research assistants were trained in the orthographic system developed by Ayafor (2014), but the spelling system necessarily had to be fully developed alongside the transcription, which required constant revision of the transcriptions as the spelling guide was developed.

5. Tagging

In this section, we describe the process for devising a tagset for CPE that is adapted from CLAWS5 (§5.1), the process of manually tagging a 10,000-word set of training data (§5.2), the initial results of automatic tagging tests using the Tree Tagger (§5.3), and the key challenges encountered during the tagging phase of the project (§5.4).

5.1. Devising a tagset

The tagset for this corpus was based on the Constituent Likelihood Automatic Word-tagging System (CLAWS) series of tagsets (Garside 1987:30), which was adapted to the structure of CPE. The CLAWS tagsets, which were used for the British National Corpus (Leech, Garside and Bryant 1994), were chosen as a starting point primarily so as to ensure comparability with other existing corpora of post-colonial varieties of English (e.g. the ICE subcorpora). The CLAWS5 tagset was selected in preference over the more recent CLAWS7 because the latter is considerably larger and, given the relative morphological simplicity of CPE, provided a more fine-grained tool than was necessary for the present project.

The typological differences between the grammar of CPE and the grammar of English meant that any attempt to 'fit' the language to existing tagsets would create an inaccurately tagged corpus. As a result, the tagset adapted for CPE differs considerably from the CLAWS5 set in a number of ways. The differences between CPE and English allowed us to reduce the number of tags for some parts of speech (POS) but required the creation of new tags for others. For example, the lack of verbal inflection in CPE contributed to a reduction in the number of tags required for lexical verbs, which in CPE have a single form and thus only require a single tag. In contrast, a category not found in English (and therefore not reflected in the CLAWS tagset) is the pre-verbal particle, which in CPE marks tense, aspect, mood, modality and negation. These were each given a unique tag in our set to allow them to be investigated individually. In addition, some features of CPE required multiple tags to reflect differences between the acrolect and the basilect. For example, nouns do not inflect for number in basilectal CPE: instead the plural particle *dem* follows a noun to indicate the plural. Speakers tending towards the acrolect do sometimes inflect for number by adding plural inflection '*–s'* to nouns (sometimes this co-occurs with the plural marker), and some nouns only occur in plural form (e.g. *dros* (< Eng. *drawers*) 'underpants'). This means that, for example, the plural noun 'books' can be expressed in CPE as either *buk dem* or *buks* or *buks dem*. As a consequence, it was necessary to include a plural noun tag in our tagset in addition to a plural particle tag.

Our analysis of the POS tags required by CPE resulted in an initial set of 42 tags. This tagset expanded as the need for additional tags in some categories became clear during the manual tagging process. Additions at this stage included individual tags for cardinal and ordinal numbers as well as tags for different categories of indefinite pronouns. The current tagset

contains 52 tags (Appendix 3). Each tag consists of three characters and retains the mnemonic significance that is a feature of the CLAWS tagsets (Garside 1987:30).

This stage of the project also led to small changes in the orthographic representation of certain words. In particular, the decision was taken at this stage to compound pronouns such as *ol ting* 'everything' and *som man* 'somebody', which show the distribution of single words. Although these had initially been transcribed as two separate words, this decision allowed a single tag to be assigned to each pronoun.

5.2. Manual tagging

Tagging a section of the corpus data manually was necessary for a number of reasons. As well as allowing the tagset to be tested and changes to be made where necessary, this process created the tagged input for training an automatic tagger. Given the time constraints of the project, it was also useful to determine the length of time required to tag the texts manually if automatic tagging were to prove unsuccessful. Three texts were chosen for the training data: two monologues and a dialogue, recorded in three different areas of Cameroon. Each text contains just over 3,000 words, resulting in 10,000 words of tagged data.

The process of manual tagging required two stages: pre-tagging, involving the preparation of the texts, followed by the tagging stage, in which each word was assigned a POS tag. The pre-tagging stage was required in part because of the changes in orthographic representation that were made while developing the tagset. This stage also allowed us to correct any irregularities in spelling and formatting, which are an inevitable feature of manually transcribed texts.

The manual tagging stage consisted of two phases: (i) group tagging of frequent words with unambiguous tags, (ii) word-by-word tagging of the remaining text.

5.2.1. Group tagging

This technique was used for unambiguous, frequently occurring words with only one possible tag. Tagging these words as a group rather than one by one (using the 'search and replace' function in a word processor) sped up the manual tagging process considerably. In part this is due to frequency; examples of particularly frequent words with a single tag include first and

third person singular pronouns *a* and *i*, which occur 586 and 382 times respectively in the 10,000 words of training data. The technique was also used on some relatively lower frequency words such as *pikin* 'child', which occurs 92 times. The usefulness of this technique is limited, as many of the most frequent words in our corpus are multifunctional (§5.4) and so cannot be tagged using this method.

5.2.2. Word-by-word tagging

This process involved tagging the remaining words in the text one by one using grammatical context as a guide for ambiguous cases. This was done longitudinally rather than cross-sectionally, to further allow context to be taken into account. The criteria for the decisions made in these instances were recorded in a tagging manual to ensure consistency. This practice allowed subsequent examples of these words and phrases to be tagged quickly. The manual tagging process also allowed us to expand the list of words that could be tagged during the group tagging stage, which, together with increasing familiarity with the language and the tagset, also increased the speed at which texts could be tagged. Manual tagging speeds increased from an average of 136 words per hour for the first text, to 185 words per hour and 300 words per hour for the second and third texts, respectively.

The output of manual tagging was 10,000 words of tagged training data, a guide to manual tagging and ambiguous cases, and a lexicon consisting of all words occurring in the tagged texts with their possible tags.

5.3. Automatic tagging test

This pilot project originally set out to tag 120,000 words, making automatic tagging desirable for this stage and essential for future larger-scale projects. As this is the first time that a corpus of CPE has been compiled and tagged, there are currently no automatic taggers for the language. It was therefore necessary to find a trainable tagger. Tree Tagger (Schmid 1994) was selected to test the possibility of automatic tagging: it is a well-established tagger, and has already been used to tag some ICE sub corpora (e.g. ICE-NIGERIA, ICE-MALTA). Tree Tagger is also readily available, and the relative ease with which it can be installed met the time constraints of this project, allowing enough time for testing, further training and tagging.

The Tree Tagger was trained and tested using manually tagged CPE data. The first training session was based on a small (6,500-word) tagged training file. The second was based on the

full 10,000 words. The same texts used to train the tagger were used to test it. The initial test was positive, with an accuracy rate of 89.3%. The second test using 10,000 words had an accuracy rate of 90.8%. Figure 2 shows the accuracy rate of the most common tags in the corpus. These results are particularly encouraging given the high level of multifunctionality in CPE.

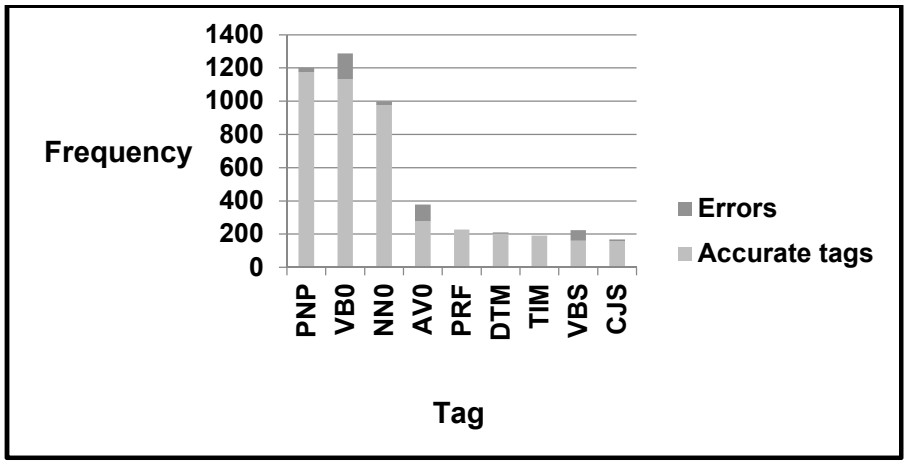


Figure 2. Accuracy of the nine most common tags in the CPE corpus

In both tests, the word most often tagged incorrectly was *goe*. This result is not unexpected: as well as being the lexical verb 'go', the form *goe* can also function as a serial verb and a preverbal particle of irrealis mood, both of which have very similar distributions in CPE. In both tests *goe* was tagged incorrectly 47% of the time and the tagger made errors with all three applicable tags, showing no improvement from the added training data.

After a final round of training, Tree Tagger has reached a 94% accuracy rate. The significant increase in accuracy rate allowed for rapid progress, and allowed for all 240,000 words to be tagged, 60,000 of which have also been post-checked.

5.4. Tagging challenges

The challenges involved in the tagging process resulted both from the features of the language and from the constraints of the project.

With respect to the features of the language, a particular challenge of tagging CPE is that the different parts of speech that a multifunctional form belongs to can often be closely related (e.g. lexical verb *goe* and serial verb *goe*). This makes it necessary to look not only at the

grammatical context of an expression but also at the semantic/pragmatic context in order to decide the correct tag. This naturally affects both the speed of manual tagging as well as the accuracy of automatic tagging, meaning that automatically tagged texts require a considerable amount of manual post-checking.

In addition, CPE can vary a great deal even within the speech of a single individual, as well as across speakers. Anglophone speakers of CPE often alternate between points on the CPE-English continuum, and judgement calls often have to be made on where the dividing line between CPE and English should be drawn.

Further challenges associated with tagging a spoken corpus include features such as hesitation, repetition and interrupted speech, which often result in ungrammatical strings and can present challenges to the tagger, human or automatic. These difficulties can be resolved to some extent by writing rules to remove disfluent strings from tagger input, thus preserving underlying transitional probabilities in the data. Naturally, there is a danger that overuse of this approach can lead to an idealised corpus which, while easier to tag, does not reflect the reality of spoken CPE.

The time constraints involved in tagging a pilot corpus present a further challenge. The number of words that could feasibly be transcribed and tagged within the timeframe of our project was necessarily limited, thus the quantity of potential training data and the accuracy rate of automatic tagging were similarly limited.

A final issue is the need for every stage of the tagging process to remain flexible: it is important that changes can continue to be made to the tagset if they are required.

Other factors requiring ongoing revision include decisions about what should be contained in the mark-up, what can be considered to be CPE (as opposed to codeswitching into English or adstrate languages), and sometimes even which part of speech corresponds to a particular word. To this end we endeavoured to make the tagging system as adaptable as possible so that it could be revised as new transcribed texts became available during the course of the project.

6. Case studies

In order to evaluate the comparability of the CPE corpus both to (i) the ICE project, and (ii) to other corpora of varieties of English in Africa, we conducted some (lexical) case studies comparing the CPE corpus with ICE-NIGERIA (Wunder et al. 2010) and with the Corpus of Cameroon English (CCE), consisting of over 800,000 words of written standard Cameroon English (Tiomajou 1993; Nkemleke and Mbangwana 2007).

To investigate the hypothesis that there is a strong relation between West African Pidgin English (WAPE) and their corresponding national variety of West African (Standard) English (WAE), Peter and Wolf (2007) compare Ghanaian Pidgin English, Nigerian Pidgin English, and Cameroon Pidgin English with their corresponding regional standards (i.e. Ghanaian English, Nigerian English, and Cameroon English). They find that the structural features of WAPEs and WAEs correspond quite closely, especially in phonology, less so in lexis. These authors conclude that WAPE and WAE varieties are (structurally) not independent of one another (2007:18). In light of these findings, we would expect similarities between findings derived from our CPE corpus and from CCE.

6.1 Lexical frequency

In these three corpora (CPE, ICE-NIGERIA, CCE), the ratio of lexical-to-grammatical words in the 40 most frequent words immediately shows typological differences between the languages: CPE has a high proportion of multifunctional items (e.g. *foe* (< Eng. *for*), which can be a preposition and an infinitival marker; or verbs such as *meik* 'make', which can function as a lexical verb, as a serial verb, or as a modality marker). CPE also lacks inflected forms. Both of these typological features allow more space in the top forty for lexical items.

CPE (spoken)	**CCE (written)**	**ICE-NIG (written)**
0.4	0.03	0.04

Table 2: lexical-to-grammatical word ratios in CPE, CCE and ICE_NIGERIA

A closer look at the frequency lists confirms this (Appendix 4): the CPE list is (predictably) heavily populated by grammatical elements such as determiners, pronouns and tense/aspect/mood/modality markers, but verbs of general meaning (*wan* 'want', *tok* 'say', *si*

'see', *nou* 'know') also make an appearance in the 40 most frequent lexemes. This is not the case for CCE or ICE-NIGERIA. Specifically, the only lexical word that makes the top 40 in CCE is *Cameroon*, which can be explained by looking at the nature of the corpus (written texts from legal and journalistic sources). On the other hand, in ICE-NIGERIA, the only word without a grammatical function in the top 40 is the verb *know* (a general meaning verb, much like those in the CPE list).

In CPE, the most frequent word is *foe*, a multipurpose preposition and also an infinitival particle. This is by far the most frequent preposition in CPE, whereas both CCE and ICE-NIGERIA have a number of prepositions in the top 40 (*to*, *of*, *in*, *for*, *with*, *on*, *from*, *at*).

In the CPE corpus, verbal elements with both lexical and grammatical uses (e.g. *goe*, which can function as lexical verb, as a serial verb, and as a preverbal particle of irrealis mood) are well represented. On the other hand, verbs of general meaning (GET, MAKE, and TAKE) are used much more frequently in CPE (ranking 25, 24 and 51 in frequency, respectively) than in both CCE (ranking 113, 47, 60, respectively) and ICE-NIGERIA (ranking 76, 73, 75, respectively).

CCE seems to pattern similarly to ICE-NIGERIA, i.e. multiple forms of BE and HAVE appear in the top 40, a not unexpected occurrence since they can convey both lexical and grammatical meanings (i.e. both verbs function as lexical as well as auxiliary verbs expressing aspectual distinctions).

6.2 GIVE ditransitives

According to Schröder (2013), in the case of GIVE ditransitives, CPE speakers favour the indirect-object construction (DAT) 70% of the time over the double-object construction (DOC), which makes up the remaining 30%. In other words, structures such as (1) would be dispreferred in favour of (2):[4]

(1) a don bai yu som buk dem
1S PF buy 2S INDEF book PL
'I've bought you some books.'

[4] Abbreviations: 1S = first person singular pronoun; 2S = second person singular pronoun; INDEF = indefinite determiner; PF = perfective aspect marker; PL = nominal plural marker; PREP = preposition;

(2) a don bai som buk dem foe yu
1S PF buy INDEF book PL PREP 2S
'I've bought some books for you.'

Searches in our CPE corpus, however, fail to confirm Schröder's observations. In fact, the converse holds: for GIVE ditransitives, DOC is the preferred pattern (74%) over DAT (26%).

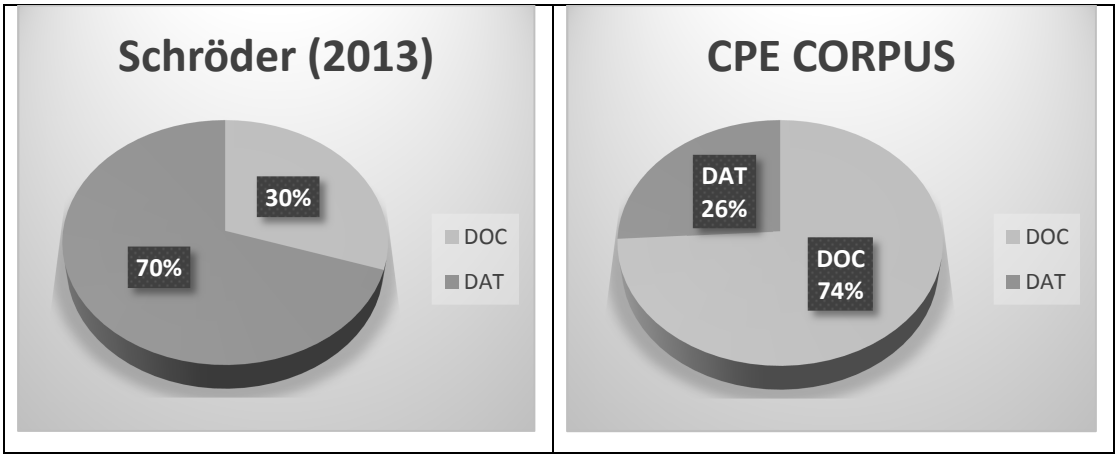


Figure 3: GIVE ditransitives in CPE

This result from our corpus ties in with the observation made by Michaelis (2014) that the DOC strategy is favoured by African pidgin/creole languages, and the DAT strategy by their European lexifiers. Furthermore, a quick measure of comparison reveals that CPE speakers' preference for the DOC pattern mirrors that of speakers in ICE-NIGERIA.

	CPE (spoken)	**ICE-NIG (spoken)**
DOC	74%	68%
DAT	26%	32%

Table 3: GIVE ditransitives in spoken Cameroon Pidgin English and spoken Nigerian English

This trend is reversed (with DAT being the favoured strategy) when we look at written corpora: CCE results patterns quite closely with the written component of ICE-NIGERIA.

	CCE (written)	ICE-NIG (written)
DOC	44%	34%
DAT	56%	66%

Table 4: GIVE ditransitives in written Cameroon English and written Nigerian English

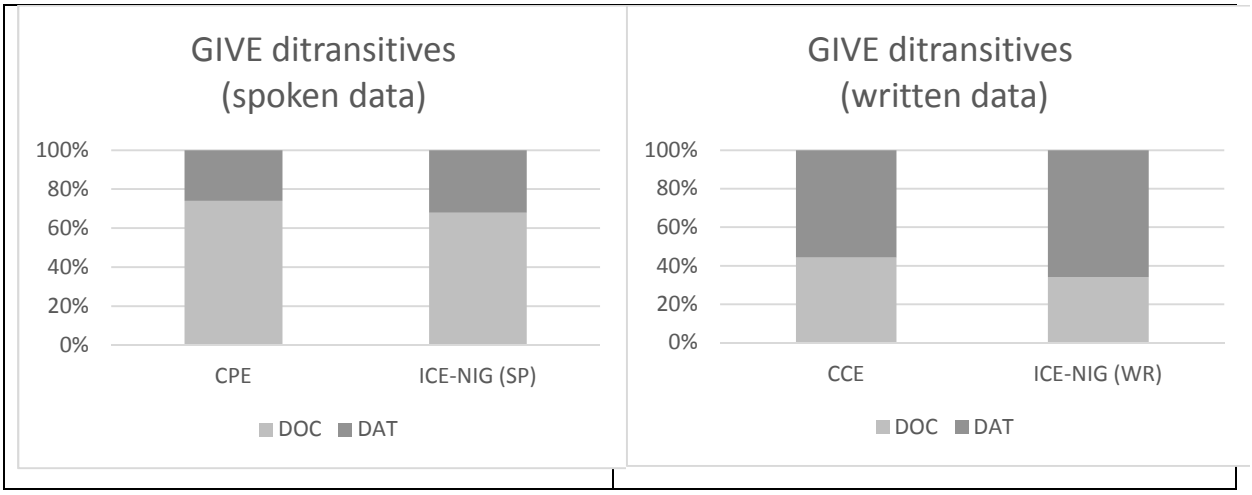


Figure 4: GIVE ditransitives in spoken/written corpora

These comparisons demonstrate the comparablility of the corpora, allowing the identification of dimensions (in this case, spoken vs. written) that exert similar effects, regardless of the language type in question. In other words, the patterns found correlate not with data source (i.e. whether they come from a particular regional variety, with varying degrees of codification), but rather with mode of communication.

7. Conclusions and prospects

Besides operational and other expected difficulties in design (representativeness and comparability) and compilation (collection, transcription, annotation), a corpus of a non-standard spoken variety poses certain challenges of its own. Achieving balance and representativeness is a particular challenge in such a complex multilingual environment, where even the speech of an individual may display considerable lectal variation. In addition, despite its widespread use, CPE lacks a standard written form, entailing that an appropriate and properly motivated spelling system had to be developed prior to the transcription stage.

Furthermore, a designated tagset with sufficient granularity had to be developed for the language, taking into account the typological differences between English and this pidgin/creole variety. This task faced the additional challenge that comprehensive grammatical description was at a relatively early stage. Despite these challenges, automatic tagging tests have provided promising results, with a 94% accuracy rate.

The case studies described above (§6) offer a snapshot of the potential uses of the CPE corpus for researchers interested in comparative studies of English worldwide. Much like other second-language ICE corpora, the CPE corpus represents the first attempt to provide a systematic database of a not yet codified variety emerging from a highly complex contact situation. The CPE corpus can be expected to significantly expand the database illustrating the spectrum of specific and general variation found in pidgin/creole varieties and regional standards, which has been identified as an exciting new research avenue (Peter and Wolf 2007; Deumert 2010; Hackert 2010).

On completion of the pilot project in summer 2016, the sound files and texts were deposited with the Oxford Text Archive. Funding is being sought for the compilation of a larger 1,000,000-word corpus of spoken CPE, which would allow more robust linguistic generalisations to emerge. More substantial funding would also allow us to address the limitations identified by the pilot project, with a view to increasing representativeness, comparability and balance of the corpus.

References

Anchimbe, Eric A. (ed.). 2012. *Language contact in a postcolonial setting: the linguistic and social context of English and Pidgin in Cameroon.* Berlin: Mouton de Gruyter.

Ayafor, Miriam. 1996. An orthography for Kamtok. *English Today* 12: 53–57.

Ayafor, Miriam. 2004. Cameroon Pidgin English (Kamtok): Morphology and Syntax. In Kortmann, Bernd, Edgar W. Schneider, Kate Burridge, Rajend Mesthrie and Clive Upton (eds.) *A Handbook of Varieties of English*. Vol. 2. Berlin: Mouton de Gruyter. 909–928.

Ayafor, Miriam. 2014. Cameroon Pidgin English orthography. Paper presented at the 14th International Colloquium of Creole Studies, Aix-en-Provence, 29-31 October 2014.

Ayafor, Miriam and Melanie Green (in press). *Cameroon Pidgin English* [London Oriental and African Language Library]. Amsterdam: John Benjamins.

Ayafor, Miriam, Melanie Green and Gabriel Ozón. In preparation. *A spoken corpus of Cameroon Pidgin English: a pilot study*. Ms, University of Sussex.

Ayisi, Florence and Kim Longinotto (dir.) 2005. *Sisters in Law* [DVD]. London: Vixen Films.

Bellama, David, Solomon Nkwelle and Joseph Yudom. 2006. *An introduction to Cameroonian Pidgin*. 3rd ed. Cameroon Peace Corps.

Biewer, Carolin, Marianne Hundt and Lena Zipp. 2010. 'How' a Fiji corpus? Challenges in the compilation of an ESL ICE component. *ICAME Journal* 34: 5–23.

de Féral, Carole. 1989. *Pidgin-English du Cameroun. Description linguistique et sociolinguistique*. Paris: Peeters/Selaf.

Dwyer, David. 1966. *An introduction to West African Pidgin English*. Michigan State University: African Studies Center.

Deuber, Dagmar. 2010. Standard English and situational variation: Sociolinguistic considerations in the compilation of ICE-Trinidad and Tobago. *ICAME Journal* 34: 24–40.

Garside, Roger (1987). The CLAWS Word-tagging System. In R. Garside, G. Leech and G. Sampson (eds.), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Gilman, Charles. 1972. *The comparative structure in French, English, and Cameroonian Pidgin English: an exercise in linguistic comparison*. PhD Thesis, Northwestern University.

Greenbaum, Sidney. 1991. *The compilation of the International Corpus of English and its components*. London: Survey of English Usage.

Greenbaum, Sidney (ed.). 1996. *Comparing English worldwide: The International Corpus of English.* Oxford: Clarendon.

Hackert, Stephanie. 2010. ICE Bahamas: Why and how? *ICAME Journal* 34: 41–53.

Holm, John (1989) *Pidgins and Creoles. Vol. 2: Reference Survey*. Cambridge: Cambridge University Press.

Huber, Magnus (1999) *Ghanaian Pidgin English in Its West African Context: A Sociohistorical and Structural Analysis*. Amsterdam: Benjamins.

Koenig, Edna L., Emmanuel Chia and John Povey (eds.). 1983. *A sociolinguistic profile of urban centres in Cameroon*. Los Angeles: Crossroads Press.

Kouega, Jean-Paul. 2008. *A dictionary of Cameroon Pidgin English Usage*. Munich: Lincom.

Leech, Geoffrey, Roger Garside and Michael Bryant (1994). CLAWS4:The tagging of the British National Corpus. In *Proceedings of the 15th International Conference on Computational linguistics (COLING 94)* Kyoto, Japan. Accessed online at http://ucrel.lancs.ac.uk/claws/.

Lewis, M. Paul, Gary F. Simons and Charles D. Fennig (eds.). 2014. *Ethnologue: language of the world*. 17th edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Mbangwana, Paul N. 1983. The scope and role of Pidgin English in Cameroon. In Koenig, Edna L., Emmanuel Chia and John Povey (eds.) *A sociolinguistic profile of urban centres in Cameroon*. Los Angeles: Crossroads Press.79–91.

Mbassi-Manga, Francis. 1973. *English in Cameroon: a study of historical contacts, patterns of usage and common trends*. Unpublished PhD dissertation, University of Leeds.

Menang, Thaddeus. 2004. Cameroon Pidgin English (Kamtok): Phonology. In Kortmann, Bernd, Edgar W. Schneider, Kate Burridge, Rajend Mesthrie and Clive Upton (eds.) *A Handbook of Varieties of English*. Vol. 1. Berlin: Mouton de Gruyter. 902–917.

Michaelis, Susanne. 2014. *Loan valency patterns in creoles: Evidence from the Atlas of Pidgin and Creole Language Structures (APiCS)*. Paper presented at the 47th Annual Meeting of the Societas Linguistica Europaea, Adam Mickiewicz University, Poznań, Poland.

Mukherjee, Joybrato, Marco Schilk and Tobias Bernaisch. 2010. Compiling the Sri Lankan component of ICE: Principles, problems, prospects. *ICAME Journal* 34: 64–77.

Nelson, Gerald. 1996. The design of the corpus. In Sidney Greenbaum (ed.). *Comparing English worldwide. The International Corpus of English.* Oxford: Clarendon Press, 27–35.

Nelson, Gerald. 2002. *Markup manual for spoken texts*. <http://ice-corpora.net/ice/spoken.doc>

Ngefac, Aloysius. 2014. The evolutionary trajectory of Cameroonian Creole and its varying sociolinguistic statuses. In S. Buschfeld, T. Hoffmann, Magnus Huber and A. Kautzsch (eds.) *The evolution of Englishes*. Amsterdam: John Benjamins. 434–447.

Nkemleke, Daniel and Paul Mbangwana. 2007. *Manual of information to accompany the Corpus of Cameroonian English (CCE)*. Department of English, Chemnitz University of Technology, Germany.

Nkengasong, J. Nkemngong. 2016. *A grammar of Cameroonian Pidgin*. Newcastle upon Tyne: Cambridge Scholars.

Peter, Lothar and Hans-Georg Wolf. 2007. A comparison of the varieties of West African Pidgin English. *World Englishes* 6.1: 3–21.

Sala, Bonaventure M. 2014. Writing in Cameroon Pidgin English: begging the question. *English Today* 25: 11–17.

Schmid, Helmut (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK. Accessed online at ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf

Schneider, Gilbert D. 1966. *West African Pidgin-English: a descriptive linguistic analysis with texts and glossary from the Cameroon area.* Hartford Seminary Foundation.

Schröder, Anne. 2003. *Status, functions and prospects of pidgin English: an empirical approach to language dynamics in Cameroon*. Tübingen: Gunter Narr Verlag.

Schröder, Anne. 2013. Cameroon Pidgin English structure dataset. In Michaelis, Susanne Maria, Maurer, Philippe, Haspelmath, Martin and Huber, Magnus (eds.) *Atlas of Pidgin and Creole Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://apics-online.info/contributions/18, accessed on 2016-05-07.)

Simo Bobda, Augustin and Hans-Georg Wolf. 2003. *Pidgin English in Cameroon in the New Millenium*. In Lucko, Peter, Peter Lothar and Hans-Georg Wolf (eds.). Studies in African Varieties of English. Frankfurt: Peter Lang. 101–117.

Tiomajou, David. 1993. Designing the corpus of Cameroon English. *ICAME Journal 17*: 119–124.

Todd, Loreto. 1969. *Pidgin English of West Cameroon*. PhD thesis: Queen's University Belfast.

Todd, Loreto. 1982. *Cameroon.* Heidelberg: Julius Groos Verlag.

Todd, Loreto. 1991. *Talk Pidgin. A structured course in West African Pidgin English.* Leeds: Tortoise Books.

Wolf, Hans-Georg. 2001. *English in Cameroon*. Berlin: Mouton de Gruyter.

Wunder, Eva-Maria, Agilantis Holger Voormann and Ulrike Gut. 2010. The ICE Nigeria corpus project: Creating an open, rich and accurate corpus. *ICAME Journal 34*: 78.88.

Appendix 1: CPE orthography

IPA	Description	Orthographic symbol	Example	Gloss
/a/	low central unrounded	a	/kam/ kam	‘come’
/ /	mid-low front unrounded	e	/g t/ get	‘get’
/e/	mid-high front unrounded	ei	/tek/ teik	‘take’
/i/	high front unrounded	i	/si/ si	‘see’
/u/	high back rounded	u	/luk/ luk	‘look’
/ /	mid-low back rounded	o	/l k/ lok	‘lock’
/o/	mid-high back rounded	oe	/go/ goe /got/ gote	‘go’ ‘goat’

Table 1: CPE vowels

IPA	Orthographic symbol	Example	Gloss
/ai/	ai	/bai/ bai	‘buy’
/au/	au	/kau/ kau	‘cow’
/ i/	oi	/b i/ boi	‘boy’
/ia/	ia	/bia/ bia	‘beer’
/i /	ie	/hi / hie	‘hear’

Table 2: CPE diphthongs

IPA	Description	Grapheme	Example	Gloss
/p/	voiceless bilabial stop	p	/**p**ut/ put	‘put’
/b/	voiced bilabial stop	b	/**b**il/ bil	‘build’
/t/	voiceless alveolar stop	t	/**t**ek/ teik	‘take’
/d/	voiced alveolar stop	d	/**d**ans/ dans	‘dance’
/k/	voiceless velar stop	k	/**k**au/ kau	‘cow’
/g/	voiced velar stop	g	/**g**if/ gif	‘give’
/t /	voiceless palatal-alveolar affricate	ch	/**t** p/ chop	‘eat’

/d /	voiced palatal-alveolar affricate	j	/ / joj	‘judge’
/f/	voiceless labiodental fricative	f	/**f**aif/ faif	‘five’
/v/	voiced labiodental fricative	v	/**v**ois/ vois	‘voice’
/s/	voiceless alveolar fricative	s	/**s**abi/ sabi	‘know’
/z/	voiced alveolar fricative	z	/i**z**i/ izi	‘easy’
/ /	voiceless palatal-alveolar fricative	sh	/ us/ shus	‘shoes’
/h/	voiceless glottal fricative	h	/**h**au/ hau	‘how’
/m/	bilabial nasal	m	/**m** k/ mek	‘make’
/n/	alveolar nasal	n	/**n** m/ nem	‘name’
/ŋ/	velar nasal	ng	/tr **ŋ**/ trong	‘strong’
/l/	alveolar liquid	l	/**l**uk/ luk	‘look’
/r/	alveolar trill	r	/**r**ot/ rot	‘road’
/w/	bilabial glide	w	/**w** / wosh	‘wash’
/j/	palatal glide	y	/**j**i / yie	‘year’

Table 3: CPE consonants

Appendix 2: Mark-up symbols

Symbol	Function
<#>	text unit marker
<$A>	speaker identification
<+> </+>	overlapping speech
<ant> </ant>	anthropophonics
<foreign> </foreign>	foreign word(s)
<indig> </indig>	indigenous word(s)
<@> </@>	changed name or word
<unclear> </unclear>	unclear word(s)
<?> </?>	uncertain transcription
<O> </O>	untranscribed text
<&> </&>	editorial comments

Appendix 3: CPE pilot tagset

AJ0	adjective
AV0	adverb
AVE	emphatic adverbial
AVQ	interrogative adverb
CJC	coordinating conjunction
CJS	subordinating conjunction
CMC	complementiser
CMR	relativiser
CNI	non verbal identificational copula
CNL	locative copula
CV0	verbal copula
DGC	cardinal numeral
DGO	ordinal number
DTA	article
DTD	demonstrative determiner
DTM	possessive determiner
DTN	quantificational determiner
DTQ	interrogative determiner
FOC	focus marker
FOR	foreign word
IDG	indigenous word
INA	infinitive marker (acrolectal)
INF	infinitival particle
ITJ	interjection
NEG	negative particle
NEP	negative perfective particle
NN0	common noun
NN2	plural noun (acrotlectal)
NNP	proper noun
NPL	plural particle
PIA	indefinite pronoun: assertive existential

PIE	indefinite pronoun: elective existential
PIN	indefinite pronoun: negative
PIU	indefinite pronoun: universal
PIW	indefinite pronoun: specific
PND	demonstrative pronoun
PNM	possessive pronoun
PNP	personal pronoun
PNQ	interrogative pronoun
PNR	reflexive/reciprocal pronoun
PRF	preposition
PRP	preposition other
PUN	punctuation marker
RES	resumptive element
SDP	serial deontic particle
TAN	anterior marker
TIM	imperfective aspect marker
TIR	irrealis marker
TMO	modality marker
TPE	perfective aspect marker
VB0	lexical verb
VBS	serial verb

Appendix 4: Frequency list/s

Rank	CPE pilot	ICE-NIGERIA (spoken)	CCE (written)
1	*foe* (preposition; infinitive particle)	*the*	*the*
2	*i* (third person singular subject pronoun)	*to*	*of*
3	*a* (first person singular subject pronoun)	*you*	*to*
4	*di* (imperfective aspect marker)	*that*	*and*
5	*yu* (second person singular subject pronoun)	*of*	*in*
6	*goe* (lexical/serial 'go', irrealis mood marker)	*and*	*and*
7	*na* (non-verbal copula, focus marker)	*i*	*is*
8	*wei* (relativiser)	*erm*	*that*
9	*sei* ('say', complementiser)	*is*	*for*
10	*am* (third person singular/plural clitic pronoun)	*in*	*be*
11	*de* (definite determiner)	*it*	*it*
12	*noe* (negation marker)	*a*	*he*
13	*dat* (distal demonstrative determiner/pronoun)	*we*	*as*
14	*dem* (third person plural object/topic/focus pronoun)	*s*	*i*
15	*deiy* (locative copula/adverb)	*are*	*you*
16	*soe* (conjunction 'so'/adverb 'thus')	*have*	*are*
17	*dey* (third person plural subject pronoun)	*this*	*with*
18	*bi* (copula/anterior tense marker)	*so*	*on*
19	*wi* (first person plural pronoun)	*for*	*not*
20	*wan* ('want')	*they*	*this*
21	*don* (perfective aspect marker)	*not*	*was*
22	*mi* (first person singular object/topic/focus pronoun)	*know*	*by*
23	*yi* (third person singular object/topic/focus pronoun)	*be*	*have*
24	*meik* (lexical/serial/modal 'make')	*he*	*his*

25	*get* (‘have’)	*on*	*from*
26	*kam* (lexical/serial ‘come’)	*was*	*at*
27	*dis* (proximal demonstrative determiner/pronoun)	*will*	*they*
28	*man* (‘man, person’)	*but*	*will*
29	*ting* (‘thing’)	*what*	*or*
30	*ma* (modal particle)	*t (-n't)*	*which*
31	*tok* (‘say’)	*one*	*all*
32	*som* (indefinite determiner/pronoun)	*as*	*we*
33	*tu* (‘two, too’)	*there*	*s*
34	*nau* (‘now’)	*now*	*cameroon*
35	*an* (co-ordinating conjunction)	*mhm*	*has*
36	*taim* (‘time’)	*okay*	*who*
37	*eh* (interjection)	*if*	*their*
38	*si* (‘see’)	*yes*	*but*
39	*nou* (‘know’)	*with*	*one*
40	*fit* (modal particle)	*at*	*an*