DOI: 10.1111/weng.12350

WILEY WORLD ENGLISHES

PAPER

The International Corpus of English project: A progress report

John Kirk[1] | Gerald Nelson[2]

[1]Institute of English and American Studies, University of Vienna

[2]The Chinese University of Hong Kong, Department of English

Correspondence
John Kirk, Institute of English and American Studies, University of Vienna, Spitalgasse 2, 1090 Wien, Austria.
Email: john.kirk@univie.ac.at

Abstract

This article begins by introducing the International Corpus of English project and proceeds to summarize the findings and outcomes of an extensive review by written questionnaire conducted by the present authors (Kirk & Nelson, 2017). Although critically concerned with practice hitherto, the review also discusses possible second generation components, and the issues in need of addressing before they should begin. The report contains many comments from a questionnaire that respondents complete, giving a flavour of the importance with which the corpus is valued. Respondents also raise a number of fundamental questions about the nature of L2 varieties of English in multilingual contexts. An Appendix sets out a prospectus for a possible component of electronic texts. Other Appendices list the corpus's text categories and their quantities as well as the 27 national components and their directors.

1 | INTRODUCTION

The study of world Englishes has come a long way in the past thirty years or so, ever since the establishment of the journals *English World-Wide* in 1980 and *World Language English* in 1981 and the impetus of the regional overviews collected in Bailey and Görlach (1982). From then on, the field has developed in no small measure through an explosion of national-variety data that have increasingly become available as well as the empirical research which has conducted on those data often on a comparative basis. Foremost amongst such comparative resources has been the *International Corpus of English*: a collection of written and spoken texts of the same types and amounts, which have been collected and designed intentionally, to provide a balanced representation of standardized varieties of English. The corpus currently comprises 27 national components: nine as L1/ENL varieties or ‘core Englishes’ (such as Australia, Canada, GB, and Ireland) and 18 as L2/ESL varieties or ‘new Englishes’ (such as Jamaica, Nigeria, Singapore, and the Philippines) (see Appendix 2), corresponding to Kachru's (1985) model of ‘Inner’ and ‘Outer’ Circles of English. The 200 written texts (each of 2,000 words) encompass printed and non-printed texts, with the former ranging from informational and instructional writing to persuasive and creative writing (see Appendix 1); inevitably, the written texts approximate to local varieties of standardized English. The 300 spoken texts (each again of 2,000 words) encompass 15 discourse situations, ranging from private and public dialogues to scripted and unscripted monologues; because of their situational contexts (broadcasting, law courts, education, and so on) and from the language used in them, an approximation towards spoken standardized English may also be inferred. In the light of subsequent research (most notably Schneider, 2007), not all ICE L2 (or ESL) varieties in their spoken form are standardized. All speakers are expected to be adults (over 18 years of age) and have completed their high school education – in fact, a great many speakers are graduates.

Following the publication of *A comprehensive grammar of the English language* (Quirk et al., 1985), which had been somewhat informed by the *Survey of English Usage Corpus* (the spoken part of which had been computerized as the *London-Lund Corpus of Spoken British English*) and which had shown awareness of British and American differences in the standardized language, it seemed a natural development to extend the description of the standardized language to national varieties of English worldwide, both where English is a mother tongue or native language, and those where it is an official or second or additional language. As McEnery and Hardie (2012, p. 75) comment: ‘by comparison with what went before, the *Comprehensive Grammar* provided a new model for a corpus-based grammar’. Thus, to that end, in an article in *World Englishes* exactly thirty years ago, a new corpus came to be proposed: the *International Corpus of English* (ICE) (Greenbaum, 1988). For its initiator and primum mobile, Sidney Greenbaum, the principal aim and objective of ICE was ‘to provide the resources for comparative studies of the English used in countries where it is either a majority first language (ENL) (for example, Canada and Australia) or an official additional language (ESL) (for example, India and Nigeria). In both language situations, English serves as a means of communication between those who live in these countries. The resources that ICE is providing for comparative studies are computer corpora, collections of samples of written and spoken English from each of the countries that are participating in the project’ (Greenbaum, 1996, p. 3). Following Greenbaum's initial proposal (Greenbaum, 1988), discussions were held in 1989 at the 10th international conference on computer corpora on English language research held in the name of the International Computer Archive of Modern and Medieval English (ICAME), which was held in Bergen (Johansson & Stenström, 1991), where the project was inaugurated. At the 11th ICAME conference, in Berlin, in 1990, the main details of the corpus were discussed and agreed (Leitner, 1992a). At subsequent ICAME meetings, further arrangements were made for annotation schemes.

Thirty years on, given the many successes and achievements which the corpus has facilitated, with radical changes in technology, with ICE teams having dropped a generation themselves, and moreover with a changing world with increased global travel and migration, increased literacy, education and entry to higher education, it seemed warranted to review its practices and to consider how best the project may be developed in the next 30 years. With those 27 national L1 and L2 components compiled or being compiled (Appendix 2), ICE is truly a worldwide project. As the review came to show, however, complications inevitably arise from the need for conformity to and replication of a prearranged plan, such as the different legal and cultural contexts in which data are collected and treated, to which some cognizance has had to be given. Copyright law is not uniform, and some countries are more restrictive than others, leading to different solutions. Financial, even material resources for corpus compilation differ from country to country, and some countries could only be included because funding has come from European universities. Attitudes towards local as well as international co-operation differ from country to country, making data collection far from easy or uniform. With teams having started at different times and are at various stages of completion, some components are no longer contemporaneous with others. Despite an agreed set of protocols for collection, transcription, markup and annotation, which most teams have endeavoured to follow, a few teams have sought to go their own way. Although, as ICE Coordinator, the second named author of this paper has strenuously pursued the creation of proverbial unity amongst diversity to prevent fragmentation, he has found it increasingly challenging to create overall a prevailing, unifying ethos. As Nelson cautions, the project's single most important objective should be: ‘to ensure that the project does not fragment into separate, non-comparable regional corpus projects. As time goes on, this fragmentation becomes more and more likely, as teams increasingly disregard the original agreed protocols. This is particularly true of some new teams’. (This article includes a number of verbatim quotations from the questionnaire responses, without author attribution.)

Early critical discussion was concerned about the Britishness of the text categories and the difficulty of collection in L2 countries; which countries for inclusion; questions of sampling and social representativeness; and funding (Leitner, 1992b; Schmied, 1996). In the early days, orthographic transcriptions and the insertion of structural mark-up were undertaken with only very general guidelines, and usually only later was any annotation (for example, for part-of-speech tagging) added. In the meantime, hardware and corpus-exploitation software have been revolutionized, greatly facilitating accessibility and convenience for researchers, and further necessitating this review. All the same, the goal of creating *an* or even *the* international corpus has inspired everyone, and the success lies not just in the compatibility of a great many of the national components but in the countless theses, monographs and research articles which have been based on the ICE material hitherto, now all too copious to encapsulate within a single ICE bibliography.

The purpose of the present article is to summarise the main critical findings as well as outcomes of the review (Kirk & Nelson, 2017), thereby offering an insight into how ICE is used and valued by its main creators and users. The review was conducted during the winter of 2016–2017 by the present authors in collaboration with the main directors of the national components. It took the form of an extensive questionnaire, one part looking forward towards the development of second generation corpora, and the second part taking stock of the 27 national components completed or currently being compiled. Sixty two questionnaires were distributed among ICE teams, and 33 responses received. Our Interim Review, dated 26 April 2017, contained numerous recommendations, to which 19 responses were returned, each in broad approval and agreement. Non-responses were interpreted as approval. Such suggestions as were made were incorporated into a Final Review dated 16 May 2017 (Kirk & Nelson, 2017), which we subsequently summarized in a plenary on 25 May 2017 at the 38th ICAME conference on computer corpora on English language research, in Prague. Since then, ongoing discussions have led to the transfer of the project's new home base in Zurich. The review is further occasioned by the second named present author's decision, after many devoted years of service, to stand down as ICE Coordinator. (A tribute of appreciation to Gerald Nelson for his unfailingly supportive and unstintingly selfless leadership as Coordinator was made in Prague.) As Greenbaum's initial proposal appeared in *World Englishes*, it seems appropriate that, exactly 30 years later, a report of this review should appear in the same journal.[1]

2 | GREENBAUM'S VISION

As expressed in various articles[2] culminating in the volume which he edited in 1996, Greenbaum's vision for a collection of comparable corpora of English which would underpin the envisaged comparative studies particularly of lexico-syntax and of spoken and written registers has been amply fulfilled; the claim certainly holds more strongly for written texts than spoken. 'With the written component, there is a fair amount of confidence in the claims made by various authors', writes one questionnaire respondent, but cautions: 'With the spoken component, though, the lack of prosodic/phonetic indicators for studies on the mere lexical corpus does not particularly inspire confidence in the claims made by various people'. Greenbaum's vision contained many technical components including part-of-speech tagging and syntactic parsing as integral parts of the corpus. On the basis of the TOSCA tagger (Oostdijk, 1991), a specially dedicated ICE tagset was developed (Greenbaum, 1993; Greenbaum & Ni, 1996; Quinn & Porter, 1996) and subsequently an ICE parser (Buckley, 1996; Fang, 1996). In the end, only ICE-GB made use of those tools (released in 1998, Nelson, Wallis, & Aarts, 2002), although an attempt was made to use them to annotate ICE-Philippines (Wallis, n.d.). In addition to the annotations for ICE-GB, software for undertaking analyses of—as well as displaying and outputting the results from—the tagged and parsed ICE-GB corpus were developed as the *International Corpus of English Corpus Utility Package* (ICECUP). As a corpus exploration platform, ICECUP, now in version 4, developed by Sean Wallis,[3] is designed to make it easy for researchers to investigate especially a parsed corpus and output their results (Porter & Quinn, 1996; Wallis, Aarts, & Nelson, 2000). More recently, under an initiative set up by the second named author of this paper, as Coordinator, ten ICE corpora were POS-tagged using the CLAWS7 tagset.[4] At the same time each was semantically tagged using the UCREL semantic analysis system (USAS) (Rayson, Piao, & Archer, 2004). Those POS-tagged corpora are currently being used by researchers at the University of Leuven in a large-scale comparative study entitled 'Exploring probabilistic grammar(s) in varieties of English around the world' (Heller, Szmrecsanyi, & Grafmiller, 2017; Szmrecsanyi, Grafmiller, Heller, & Röthlisberger, 2016).

That vision of Greenbaum's included yet further types of annotation: prosodic, semantic, the marking of discourse features, and alignment of audio and video recordings to the transcriptions. Although alignment has been on the agenda since the outset, it was for a long time unclear how this would be undertaken. Of the earliest corpora, only ICE-GB has been aligned. Now that software is more readily accessible (see below), several newer national components (such as ICE-Nigeria, ICE-Scotland, ICE-Bahamas and ICE-Trinidad & Tobago) are including audio-alignment as standard annotation. Only one component has been prosodically tagged: ICE-Ireland, and the resulting corpus (known as the SPICE-Ireland corpus: '*S*ystems of *P*ragmatic annotation in the spoken component of the *ICE*-Ireland corpus' (Kirk, Kallen, Lowry, Rooney, & Mannion, 2011)) has been pragmatically tagged with respect to speech acts (after the defining notions by Searle, 1969, 1976), utterance tags (which include declarative as well as interrogative polarity-marking sentence tags, but also vocatives and discourse markers used sentence- or utterance-finally), quotatives (citations of speech attributed to another speaker and supposedly rendered verbatim or directly) and, of course, discourse markers (such as *well* or *kind of*) (Kallen & Kirk, 2012; Kirk, 2016). A further initiative has been the adoption of standoff architecture using ANNIS,[5] as developed by an independent project for ICE-Ireland (Kirk, 2017).

3 | PROJECT COORDINATION

Since its inception, the ICE project has been identified not simply with the Coordinator—initially Greenbaum, and since 2001 the second named present author—but especially with the website, managed privately by Nelson, through which user licences, corpora and manuals have been distributed and much primary descriptive information about the project is made available. As an outcome of the review and subsequent discussion, the home base has moved to Zurich, with Marianne Hundt taking over as the new Coordinator, ably supported by her colleagues Hans Martin Lehmann and Gerold Schneider, and the establishment of a new website: www.ice-corpora.uzh.ch, comprising initially the same material as before, and in a similar style, to be augmented in due course as appropriate. Professor Hundt has already provided leadership to the ICE-project by instigating what she came to call 'ICE Age 2' with reference to 'ICE corpora of New Englishes in the making' (*ICAME Journal 34*, 2010) with a focus on corpora of new Englishes still in the making such as ICE-Bahamas, ICE-Fiji, ICE-Malta, ICE-Nigeria, ICE-Sri Lanka and ICE-Trinidad & Tobago. Having been initiated within the last ten years or so, these components have adopted XML-format as standard, replacing the SGML-format originally used (Wong, Cassidy, & Peters, 2011) and are benefitting from the technological aids for transcription or data management which are now available. Symposia were held and a set of papers appeared in the *ICAME Journal 34* (April 2010). These papers already constitute a review of ICE at that stage and make innovative proposals (and hence the designation 'ICE Age 2').

The move is to be welcomed for another reason: Zurich is the home base of the ICEonline project (https://es-iceonline.uzh.ch), developed over a number of years by Hans Martin Lehmann and Gerold Schneider (Lehmann, 2015; Lehmann & Schneider, 2012). Through a web-based front-end server, ICEonline currently provides online access to nine completed and a further six partially-completed ICE-corpora, individually or collectively, in any combination. Moreover, the corpora have been homogenized and regularized for markup and have also been POS-tagged with CLAWS7 and automatically parsed using a functional dependency grammar developed by Schneider (2008). Moreover, ICEonline has synchronized with the text the biodata for its component corpora insofar as those data are available for mutual investigation, often of a sociolinguistic kind. With the component corpora in one place, ICEonline is the most fully developed version of the international corpus as a single composite and consolidated entity. What is more, by being homogenized and regularized in format and for markup, by being POS-tagged and syntactically-parsed, and by harmonizing the biodata with the data, ICEonline uniquely meets several of the objectives identified by participants in the review as among the most important. It is intended that, in due course, the ICEonline site be opened under license to the wider ICE community and, indeed, any desirous bona fide researchers. (For those components already part of ICEonline, see Appendix 2.)

Many tasks await the new team in Zurich, including the maintenance of the website; obtaining, storing and making available individual corpora and their biodata through downloading; the licensing of users and gate-keeping of access;

and support for users. At the same time, it is envisaged that the ICEonline project will continue to homogenize, regularize and generally tidy up both the text and the biodata of first generation corpora as they become completed, POS-tag and parse the corpora, make them available to the wider community, and provide support. In so doing, it will meet many of the strong desires for homogenization and regularization expressed in the review. What is envisaged is that the new arrangements will go some way towards satisfying the needs for access ('the consolidation of existing ICE resources into one unified super-corpus which can be accessed and searched at one online site') and dissemination (under license, through download from a central server).[6] The review found strong desire for regular meetings and other communications, so that it seems likely that the lead in arranging annual meetings in conjunction with ICAME conferences will be taken by the Zurich team.

4 | SECOND GENERATION CORPORA

A motivation for the review was the desire for some of the earliest-completed national components to be replicated a generation or so later, as has happened, for instance, with the so-called Brown family of corpora of written English (McEnery & Hardie, 2012, pp. 97–100), that is, 'to compile parallel components for the first generation components which will allow diachronic comparison'.[7] However, there was also a feeling that additional corpora or second generation corpora should not be at the expense of or 'as an alternative to completing and fully annotating the first generation corpora'. One respondent urges firmly that 'the first generation ICE corpora should, in the first instance, be all completed and available for comparison'. Appendix 2 provides completion dates, where known, for corpora still being compiled.

Enthusiasm for second generation (replication) corpora was indicated in many responses, such as: 'I believe it is already a good time to compile second generation components, most especially because diachronic analyses of Englishes have already become a trend in English linguistics'; 'They should be updated to allow diachronic research and for comparisons with newer ICE corpora (every 25 years would be great!)'; 'Updating the earliest components would be highly beneficial and open up many new avenues for research; it would be expedient to devise an updated corpus design first and to update the earliest components accordingly'; 'That is a fantastic idea and in every sense possible it should be structured as closely as possible as the original corpus to achieve diachronic comparability'. Another respondent comments that second generation corpora 'should be constructed so that there is as much comparability to first generation as possible. Of course, adjustments will have to be made (old text types don't exist anymore, new text types have come into being)'. And, indeed, some urge the inclusion of electronic texts and emerging text types that have increased in importance in recent years. More specifically, what ICE should do, urges one respondent, is to 'set a realistic date for a new suite of corpora, to enable recording of spoken texts to be conducted in as short as possible a time frame (say, 2021/2), so we get a Brown-like three-decade interval between the early 1990s and the next suite, for diachronic studies'. There emerged some confusion over the name 'second generation'. For the present authors, 'second generation' is intended to refer to a second, later corpus of the same national variety. As such, at present, there is, then, no second generation corpus. However, 'second generation' or 'Generation 2' was understood by some to refer to the set of 'ICE Age 2' corpora because, although first-time corpora, they are using updated technological methods including audio alignment as well as making some changes to the text categories being collected. One respondent comments: 'I think any new work now must be seen as a 'Generation 2' corpus, and there can be new rules for what constitutes a Generation 2 corpus with regard to annotation, text type, and regional-demographic-political definition. If a clear picture of ICE Generation 2 is developed, then there is no reason why there shouldn't be a Generation 2 corpus for the existing corpora'.

However reservations were expressed, too. One respondent cautions that 'Again new text categories and sources are only desirable if the standard sampling frame can still be imposed on the material'. Another thinks that 'it's only really feasible if we agree on really easily accessible electronic texts. And if there's some automated way of getting transcription of spoken language. I don't see it as feasible if we try doing the data collection and transcription the way we did it before'. As another points out, this issue is answered: 'The new corpora should follow the compilation process of the “ICE-Age 2” corpora’. Specifically, the review suggested the following components should be considered candidates for being replicated as second generation corpora: ICE-India and ICE-East Africa (because of nonconformities to the standard ICE protocol and numerous complaints); ICE-Australia (because the data were never generally released); ICE-USA (the spoken component was never completed); ICE-GB (because a second generation corpus is greatly desired). As we received no responses about ICE-Namibia or ICE-Pakistan, we are unsure of their progress or status. Before second generation corpora get off the ground, we feel that it is important for serious consideration to be given to the types of text which should be gathered in a multilingual society where English (as an L2) is only one of a number of competing official languages, so that those texts may be taken as truly representive of the status and use of English in those countries. A related issue was raised: in L2 countries, was it really ‘educated’ speech that was being targeted or the use of the acrolectal variety (that is the capturing of stable features and varieties of English use in that country)? There was also the question of how mixed codes should be handled. Such discussion about which texts categories would best suit L2 countries and, indeed, for which L2 countries second generation corpora should be undertaken might make a suitable agenda for an ICE meeting in the near future. This topic is further developed in Section 5 below.

Separate from second generation corpora, although their guidelines may come to be followed, the review yielded suggestions for additional corpora, particularly more ‘regional’ or subnational corpora, such as Wales, California, Francophone Canada, or other parts of Africa. The separate proposal by Ozón, Ayafor, Green, and FitzGerald (2017) to include Cameroon Pidgin raises a challenge to the inference that the language to be found in the spoken public as well as private ICE categories is amounting to a form of standardized English. Yet it is the local creolized/pidginized variety rather than a local standardized variety that, according to Mair (2013, p. 264), would be a better reflection of Jamaica within a ‘world system of Englishes’, and more likely to be borrowed from, than standardized Jamaican English. The issue of extending ICE corpora to Kachru's ‘Expanding Circle’ where English has no official status (as an EFL) was also tested in the questionnaire but did not receive much support, although the ‘theoretical, methodological and empirical basis the expansion of the ICE concept to the Expanding Circle’ is vigorously presented as ‘ICE Age 3’ in Edwards (2016), Edwards and Laporte (2015) and now Edwards (2017) on the basis of a study of written English in the Netherlands which replicates the ICE model for text categories. A further study, on Korean English, is by Hadikin (2014). Regarding Edwards's enthusiastic plea, the second named present author, as Coordinator, cautioned: ‘I can see no real benefit to ICE of including the Expanding Circle. It would make an already large project even larger, and would produce a lack of focus in the project as a whole. A project aimed at sampling ‘every kind of English’ would be fairly unmanageable, and would not attract funding’. Although studies of English in the Expanding Circle are increasingly showing the boundaries between Outer and Expanding Circles (ESL and EFL) becoming blurred, as argued by Saraceni (2015), one seasoned ICE participant responds: ‘No Expanding Circle, and no Lingua Franca, if you ask me’.

4.1 | Sampling periods

Whereas the sampling period for the initial set of completed corpora was held constant (1990–1994), with teams subsequently starting at different times over the next twenty years or so, as already mentioned, the sampling period is now no longer parallel or identical among first generation corpora. There are limits to retrospective collection; and there is an obvious preference for the here and now, the immediate and contemporary. By stealth, temporal variation has crept in, and ‘having exactly parallel sampling periods is no longer feasible’. One solution is to annotate the data with a timestamp and/or specify dates/period of collection in the corpus handbook. As one respondent comments: ‘for collecting/transcribing/digitizing speech, time is the challenge, since changes in spoken languages (especially colloquial registers), can happen very quickly, and at different rates in different places. If the dates of data collection are clear, then at least researchers know how to factor them in as a variable’. If second generation ICE-corpora were restricted to (say) 2020–2023, it may be that some first generation corpora still being compiled would be more closely aligned to those dates than the original collection period. For one respondent, ‘an updated ICE-GB would be more comparable with certain L2 corpora currently being compiled than the original ICE-GB’.

4.2 | Speaker education

An ICE corpus contains the speech of adults over the age of 18 with completed school education. A great many speakers turn out to be university or college graduates or students, as with ICE-Ireland. Moreover, it was felt that an ICE-corpus was a corpus of adult speech (having attained and completed school education), and not of children's speech, even if, as defended in one case, ‘they are aspiring to education’. Whereas the review showed considerable agreement for the level of education to remain constant with regard to speaker choice, not least because of its importance for sociolinguistic purposes, the question was raised whether, in L2 countries, secondary education *in English* should always be stipulated? A solution might again be more flexibility, with the details about the level of education recorded as a variable.

4.3 | Spoken and written texts

As for possible revisions to the present contents, the review showed overwhelming desire for the present set of spoken and written text categories and their quantities for second generation corpora to remain unaltered, above all to ensure comparability. However, there was also a willingness to accept that each corpus need not fill text categories for which there are no data in their country. Some completed corpora have not filled some text categories where they have had difficulty gathering certain texts (legal texts; parliamentary debates; social & business letters; and so on), and this practice is acknowledged by the review under a principle of flexibility: certain text categories may simply not be available. ‘Certain text types that are not as easily obtained in certain localities due to policies, restrictions and also language (that is the language used may not be English)’. Nevertheless, the review urges that, wherever possible, each text category should be filled with the full specified quota of texts. The issue of sampling was also raised. ‘As far as possible, we should control for stable proportions of participant words by gender, age, education etc. between text categories in the same corpus’. There is probably no perfect answer to the sampling issue, but best efforts should be attempted and explained; one solution would be the inclusion of full biodata in corpus handbooks. A further recommendation under the flexibility principle is for corpora with empty text categories, likely to be in L2 countries, to add up to two new categories, each of 10 texts if deemed characteristic of the use of English in that country. Perhaps the most major change of contents for second generation corpora is the recommendation for the inclusion of a new component of electronic texts, totalling up to 500,000 words, making the total maximum size of a second generation corpus 1.5 million words. Indeed, from her thorough comprehensive comparison between ICE and the GloWbE corpus of blogs and websites (Davies, 2013), Loureiro-Porto (2017, p. 468) concludes that ‘the future of ICE should include web registers alongside the text types included so far’; from other comparisons, similar conclusions are drawn by Mair (2015), Mukherjee (2015), Nelson (2015) and Peters (2015), in their responses to Davies and Fuchs (2015). A list of possible electronic texts is presented in Appendix 3 for discussion. As a consequence of the flexibility principle, the total number of words may come to differ from corpus to corpus; and some corpora may become larger than the present words total. Whereas one-million words is a convenient norm, but with the reality of missing texts as well as the prospect of additional texts, second generation components may come to have different total numbers of words. This should not be a problem, however, as all inter-corpus comparisons can be relativized. Besides, in line with the flexibility principle, some first generation corpora have not filled every category. As one respondent remarks: ‘Enough techniques for frequency norming do exist to overcome any length-related issues, provided that they are used sensibly, that is that the norming factor is based on a common denominator’.

4.4 | Markup and annotation

As important as the choice and amounts of written texts and spoken transcriptions are, two further properties are essential for analysis and exploitation: markup and annotation. In ICE terms, markup serves to indicate the identity of texts and speakers as well as the identification of many formal aspects of an utterance such as paragraphs and sentences, turns and utterances, overlaps, pauses, comments about paralanguage, editorial insertions, and any normalizations, if indeed, included. Moreover, different practices for copy editing, anonymization, regularization, normalization and layout exist among the present constituent corpora (Nelson, 1991, 1995, 1996). Some harmonization has been undertaken for the components included in ICEonline. By contrast, annotations are additions to texts in what from now on will be an XML-format which indicate linguistic properties that a particular item or structure might have. Annotations might indicate a part of speech, a syntactic structure or function, a semantic classification, or a pragmatic function or tone movements, and so on. The review recommends that, wherever possible, corpus texts be annotated and post-edited with regard to POS tags—either using the ICE tagset (ideally) or the by now ipso facto standard tagset, CLAWS7—as a minimum. ICEonline has made considerable advancement, offering POS tagging with CLAWS7 and the PENN treebank tagset and syntactic annotation (Lehmann & Schneider, 2012; Schneider, 2008). As one respondent remarked, the addition of 'more annotations […] will make any linguistic analysis easier'. However, there was some recognition that annotation often required time-consuming manual insertion, as happened with the prosodic and pragmatic annotations in SPICE-Ireland (Kirk, 2017). A further type of annotation concerns the synchronization of the spoken transcription with the original audio or video recording, for which considerable support and desire was expressed. The review recommended that wherever possible the audio-files be digitized and aligned. 'Digitizing sound files would be an important aim', writes one respondent. 'For me', writes another, 'ICE has a competitive edge on other corpora and resources because of its focus on spontaneous speech and spoken data in general. Digitization of sound files in order to make them available to the community should be a priority'. There is a purpose in digitization: 'The sound files should be aligned with orthographic transcription, to support research on phonological variation within individual varieties, and correlate it with sentence patterns and discourse functions'. Thus another respondent urges: 'Definitely include sound files and time-aligned transcriptions for all new corpora (and where possible for the older corpora too)'.

4.5 | Biodata

An essential constituent of a corpus is the provision of sociolinguistic information about speakers. Those biodata are best presented in a database to which transcriptions can be linked for analysis and exploitation. The review recommended that detailed biodata be made easily available in an electronic format and, wherever possible, be linked to the transcription for interactive searching and exploitation. Several respondents urged for 'more standardized sociobiographical speaker annotation in all corpora (in the corpus or possibly as standoff annotation)'; for 'more uniform handbooks that go with the individual corpora'; for 'an improvement on the documentation of corpora […] to ensure cross-component comparability and the regional character of feature differences'; for 'more detailed guidelines in relation to speaker selection, text genres to be sampled, transcription, etc.'. Handbooks (in any format of dissemination) with detailed biodata would certainly be desirable for each component. Some handbooks do exist (Kallen & Kirk, 2008, 2012) but, for each corpus, where it exists, the handbook has a different format. However, in ICEonline, the biodata are encoded (again, where they were available) and, following that model, a template for biodata guidelines should be drawn up for completing first and second generation corpora. Biodata should be collated in a uniform series of electronic databases for downloading by end-users. Good biodata will be crucial for second generation corpora for highlighting generational differences as well as for profiling language change in apparent time, as shown by Hansen (2017) using data from ICE Hong Kong. 'Looking ahead', writes one respondent, 'if information on speaker education, year of production and text type differences is available and the factors are statistically controlled for when analysing features, these parallels are not maximally important; if they are not systematically controlled for in, for example, comparisons of frequencies of features across ICE components, it is paramount that the extralinguistic characteristics are as similar as possible so that frequency differences can clearly be attributed to regional (and not educational or text-type) variation'. As Edwards (2017) shows, good biodata drawing attention to educational and social backgrounds of speakers in Expanding Circle varieties is essential for comparisons with Outer as well as Inner Circle varieties; her approach certainly provides a suitable model.

4.6 | Technology and software

As an entirely computer-based project, a major concern is the need now to revise and update the methodology (which was initially addressed by the 'ICE Age 2' initiative (Gut & Fuchs, 2017; *ICAME Journal* 34, 2010). For one respondent, we need to 'update to modern technologies and data format'. For another, we need to 'bring the format up to date, that is convert it to XML' For yet another we need to 'establish modern corpus compilation standards—some teams still use text processors (for example Microsoft Word) to type up the transcriptions of spoken material. This is the technological standard of the 1980s. Such methods are much less efficient and more error-prone than modern software for corpus compilation. We need a dialogue among ICE teams that will result in recommendations of what software and corpus compilation methods should be used because the current approach (of much diversity in such these methods) negatively impacts the comparability of the corpora'. Hopefully that dialogue will come to be held at future annual meetings and a workable consensus reached. When ICE was devised in the early 1990s, technology was in a very different state from what it is today. Whereas the original arrangements have proven fit for purpose, it has become clear that newer and more recent software would beneficially facilitate the many tasks involved in the compilation of a corpus and its maintenance and exploitation. For a start, recordings are now all made digitally and may readily be stored electronically. Some respondents urged for a system that would automate the first-stage markup and transcription process (partially, at least). Although there are others (for example PRAAT and Wavesurfer), two packages have been adopted by some 'ICE Age 2' teams: ELAN, for the alignment of the transcription (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006),[8] and Pacx as a corpus management system (Gut & Fuchs, 2017; Wunder, Voormann, & Gut, 2010).[9] The use of Pacx creates XML encoding for structural markup reliably and is particularly suitable for annotation, including the linking of socio-biographical information with the transcription. Their stand-off architectures, which enable additions to the annotation to be incorporated easily, are similar to ANNIS, into which ICE-Ireland and SPICE-Ireland have been converted (Kirk, 2017). Some urged for adoption of new computer software and techniques including the transfer of annotation from SGML to XML;[10] the development of procedures for corpus compilation and annotation software, and the development of software with filters for sociobiographical and other parameters. Also urged in this connection was a way should be found to align the biodata with the texts, so that social variables (such as age and sex) can be used as search arguments in constructing queries (Hansen, 2017). One respondent urged for 'the development of a single, convenient tool with which to search all the speaker and other situational factors across ICE-corpora, ideally online […] and a search tool to exploit the annotation for all ICE corpora taken together'. This has been done for ICE-GB and, again, in ICEonline, but in most instances, the corpus and biodata are stored as separate (usually Excel) files—a relationship which Pacx has been devised to handle (Gut & Fuchs, 2017).

5 | COMPARABILITY

The review questionnaire posed the following question*: How important is it that the ICE corpora be exactly parallel to each other in terms of text types*, *sampling periods*, *speaker education*, *and so on?* One of the strongest messages to emerge from the review was the crucial need for homogenization and regularization of all the major elements in each component: text format, markup, annotation and the biodata; another was for the consolidation of the entire project. Those not inconsiderable tasks have already been achieved through the painstaking checking and editing carried out over many years largely by Hans Martin Lehmann on those corpus texts which have hitherto come to be included in ICEonline. Nevertheless, it behoves corpora still being compiled as well as second generation corpora to comply with the guidelines and protocols as fully and as accurately as possible to ensure maximum harmonization and regularization before components ever reach the server in Zurich and incorporated into ICEonline. Completion of first generation corpora has featured as a major priority in the review (as presented above); as already mentioned, several compilers have indeed indicated that they expect to have completed their components within the next couple of years (see Appendix 2).

The point of the above question about strict parallelity was affirmed very strongly: 'the basic idea and great advantage of ICE'. Similar views were plentiful: 'they [the corpora] should be exactly parallel (that was the main point when

ICE was set up)'; that comparability should be the explicit goal'; that 'keeping the external or situational variables as stable as possible ensures the comparability of the corpora—one of the key strengths of the ICE corpora'; that 'without the core sampling frame being accessible for each regional component the whole enterprise is put in danger'; that [it is] 'very important—otherwise full comparability cannot be made'; that 'it's crucial, otherwise there is no point in the project'. 'It's very important, and should be retained as an overall objective. But we must be flexible enough to adapt to local circumstances (as we have been)'. More specifically, it was urged that: '[i]t is paramount that the extralinguistic characteristics are as similar as possible so that frequency differences can clearly be attributed to regional (and not educational or text-type) variation'. Other voices suggested that the comparability should be more 'aspirational' than exactly or relatively parallel. Rather, the corpora should be 'as parallel as possible', 'as close to a common standard as possible, with local adjustments where necessary', for 'the linguistic reality is of course different in different countries' and there arises 'a need to make compromises'. 'It is very important, insofar as it is achievable'. 'We should actively strive to maintain a measure of comparability though. They aren't exactly parallel, except a number (most?) do have the same text types'. As another comments: 'though strict uniformity will never be possible, it is an essential feature (and strong selling point) of ICE that the corpora are as parallel to each other as possible'. The parallelity rests with the text categories and their quantities, which are to be maintained in second generation corpora. However, whereas diversity and variation within text categories are design features to be encouraged, there are limits to the tolerance of departures from agreed protocols. At what point does a component no longer qualify as a component of ICE? As commented by several respondents: 'This [issue of strict parallelity] is essentially the same problem facing individual corpus compilation—stratification vs. 'balanced' corpora'. 'The compatibility across corpora has always been in conflict with the specific national sociolinguistic context/basis!' A further comment is: 'I don't feel strongly that all ICE-corpora have to be exactly identical in structure, but I like it that at least some are'. One factor about which flexibility was being agreed concerns total corpus size. By not filling non-applicable categories, by adding new categories appropriate especially to L2/ESL countries, and by adding electronic texts, variability across overall total word counts would prove inevitable.

However, following on from the discussion about second generation corpora in Section 4, deeper problems were also raised, such as ecological validity: 'What are authentic English-medium genres in multilingual English cultures? Whereas ICE text categories are well representative of language distribution and use in L1 countries, it may be that in L2 and inherently multilingual countries, for a good representation of the use of English, a different set of categories is required. The desired comparability may thus need a different basis from strict parallelism in text category choices'. What functions English serves for multi-lingual speakers (and indeed does not serve) and how those spoken written genres/registers are distributed across languages are crucial aspects of the use of English in context. How languages are actually used by multilingual speakers (with the same educational and social functions as for monolingual speakers of Inner Circle Englishes?) can then be factored into their analysis, to address questions of their parallelity/comparability. A further issue is how English functions within a speaker's multilingual repertoire and in relation to their linguistic and cultural identity, on which there is increasing research, for example, for South African English speakers by Coetzee-Van Rooy (2014). Does (or should) ICE distinguish between what Spolsky (2003) calls 'multilingual societies' and 'plurilingual speakers' (speakers of more than one language)? These are all issues which affect Expanding Circle (EFL) as well as Outer Circle (ESL) countries, and which will be key to the planning of text categories for second generation corpora. As previously mentioned, choice of text categories will inevitably determine overall corpus size.

A further question was more philosophical: 'How can ICE reckon with criticisms about how strictly delineating varieties stem from a (possibly now outmoded) view of languages and language varieties as bounded and discrete?' This issue almost certainly relates to the Kachruvian model and the Quirkian notion of a 'monochrome international standard language' with only local deviations. Saraceni (2015, p. 4) argues that the world Englishes framework is 'lagging behind' sociolinguistic developments of globalization in the twenty-first century which are better explained in terms of 'super-diversity', 'hybridity', 'translanguaging' and 'metrolingualism'. As Saraceni (2015, pp. 132–134) argues, there is a need to consider 'languages across borders (§5.3.2), English vs. Non-English (§5.3.3), language switching (§5.3.4), hybridity (§5.3.5), diversity and super-diversity (§5.4.1)'. He advocates a shift from 'world Englishes to language worlds'—a shift away from analysing varieties of English as structural sets (such as in Kortmann and Schneider 2004) or 'decontextualized linguistic systems' (Mair, 2013, p. 254) to an approach which grapples with understanding 'language borders and of how people manipulate them creatively' (Mair, 2013, p. 264). Abstracting out from this dynamic, communicative approach calls to mind Mair's (2013, p. 264) new theoretical model of a 'world system of standard and non-standard Englishes', claimed by him as 'better equipped to handle uses of English in domains beyond the post-colonial nation state' (Mair, 2013, p. 253). Exactly how second generation ICE corpora should reflect English in Outer Circle countries has thus become a much more challenging issue. Besides, in a study of 'the use of pidgins and creoles in web forums serving West African and Caribbean diasporas' (Mair, 2013, p. 253), much more is known about Outer Circle varieties, their differences from standard English no longer to be regarded as substratally deviant but rather as lexifications from local pidgins and creoles in a multilingual community where the use of English(es) is but a sociolinguistically-significant choice among competing languages (Mair, 2013, 2015). How far are speakers in an Outer Circle ICE-corpus speakers of an acrolectal variety approximating to the standardized language (as envisaged at the outset of ICE) or rather simply a local mesolectal or basilectal/pidgin variety (Deuber, 2014; Mukherjee, 2015) mindful that local standardization may not have taken place (Schneider, 2007)? Add to that the ever-increasing role of English in the Expanding Circle countries, as already acknowledged, the blurring of status between EFL and ESL countries, and the growing reality of the transnational use of English as a lingua franca, there is much to consider among compilers and other project participants and interested parties in ongoing discussions about the future of ICE, not least in the forum of an annual meeting, as recommended in the review. Whatever decisions regarding second generation corpora are finally taken, the value of small-scale, carefully structured, well annotated (especially with comprehensive biodata) corpora continued to be preferred for many research purposes—not least comparable studies of national varieties—over rapidly-compiled, anonymous, indiscriminate, web-derived mega-corpora such as GloWbE (Davies, 2013; Davies & Fuchs 2015), as earlier shown, for instance, by Leech (2007) among others and now reconfirmed by Loureiro-Porto (2017) in her comparison of ICE and GloWbE.

6 | CONCLUSION

This article has presented the main outcomes of the review of the ICE project which we undertook in 2016–2017 and begun to chart the gradual evolutionary way ahead, as the project enters its next 30-year phase, with a new home base in Zurich. The review uncovered many recommendations for second generation corpora, some of which—in respect of the actual corpus data, their markup and annotation, and also the biodata—may yet come to apply to unfinished first generation corpora as well. Other issues such as copyright and research ethics could not be touched upon in this report. The biggest challenges facing second generation corpora strike us as these: to square the desire for strict parallelity with first generation corpora with the need to sample genres of texts which appropriately and adequately represent the present uses and functions of English in the national variety in question, particularly in multilingual contexts; and to consider the inclusion of Expanding Circle varieties where the public uses and speaker profiles might more readily match Inner Circle varieties than Outer Circle varieties. The reality is that there are several competing and complementary objectives, which are not, however, exclusive and may be pursued simultaneously. Completion of first generation corpora certainly remains a high priority; but annotation and alignment as well as computer software figure as priorities, too, as do the inclusion of electronic texts, good biodata, good documentation, and the much-desired regular project meetings. As one seasoned respondent writes, 'All these objectives are important, but some are more important than others'. And another: 'All this is desirable, if and only if it does not compromise on the main objective of producing comparative descriptive studies'. Here, however, is not the place to review the very many studies and uses to which ICE corpora have been put. In the 30 years since ICE was first proposed, the project has come a remarkably long way and its progress has certainly been, by any standards, extraordinarily impressive. But much more remains to be done—and could and should be done, not least harnessing to advantage the massive technological developments in recent years. The compilation of an ICE corpus has proven a complex, demanding and responsible undertaking. The project, with currently 27 component corpora, has contributed immeasurably to research and to the advancement of knowledge on world Englishes. As Loureiro-Porto (2017, p. 448) rightly remarks: ‘the validity of ICE is wholly unquestioned’. The need for ongoing continuity, effective coordination and directive leadership has never been greater.

ACKNOWLEDGEMENTS

We are deeply grateful to each of the many respondents of our questionnaire and of our Interim Report. For this paper, we are indebted to Kingsley Bolton, Sebastian Hoffmann, Hans Martin Lehmann, Gerold Schneider and Martin Weisser for discussion and comments on previous drafts; and to two anonymous reviewers for their many helpful comments and suggestions, not all of which could be dealt with here.

NOTES

[1] Since this article was drafted, *World Englishes* 36(3) has appeared, with several papers on ICE, touching on some similar points, notably in Edwards (2017), Gut and Fuchs (2017), Kirk (2017), Loureiro-Porto (2017), with an introductory overview by Nelson (2017) himself.

[2] Early references to ICE are Greenbaum (1988, 1990a, 1990b, 1990c, 1991a, 1991b, 1992, 1992, 1993, 1994, 1996a, 1996b; Nelson, 1991, 1995, 1996, 2002a, 2002b). The journal *World Englishes* devoted special issues to the ICE project in 1996, 2004 and 2017.

[3] Retrieved from https://www.ucl.ac.uk/english-usage/staff/sean/ (11 August, 2017)

[4] Retrieved from https://ucrel.lancs.ac.uk/claws7tags.html (11 August, 2017)

[5] ANNIS stands for ‘ANNotation of Information Structure’ and is available at https://corpus-tools.org/annis/ (11 August, 2017); Zeldes, Rtiz, Lüdeling & Chiarcos (2009).

[6] One respondent urged: ‘We need to find an optimal route (or routes) for disseminating corpora, so that the huge work carried out by ICE teams in the past can be used by the corpus linguistics community. If we do not do this, we will be seen as largely irrelevant in the world of ‘big data’. A dissemination project may include parallel strands: publishing on an online platform possibly tagged and published; streamlining access to downloadable data with tools; and exemplification and publicity to motivate their use’. ICEonline will largely fulfil this desire.

[7] The Brown family includes components of British and American English from 1901, 1931, 1961, 1990s and the 2000s.

[8] ELAN stands for ‘EUDICO Linguistic Annotator’ and is available at https://tla.mpi.nl/tools/tla-tools/elan/ (11 August, 2017)

[9] PACX stands for ‘Platform for Annotated Corpora in XML’ and is available at https://pacx.sourceforge.net (11 August, 2017), which states that ‘PACX is built on Eclipse, Vex, Subversive, etc. for creating and editing transcriptions and annotations, querying, managing version controlled data, and building a shippable corpus’ (Gut & Fuchs, 2017).

[10] A converter for ICE (SGML) markup into XML markup is described in Wong et al. (2011).

REFERENCES

Bailey, R. W., & Görlach, M. (1982). *English as a world language*. Ann Arbor, MI: University of Michigan Press.

Biber, D., & Kurjian, J. (2007). Towards a taxonomy of web registers and text types. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 109–131). Amsterdam: Rodopi.

Buckley, J. (1996). An outline of the survey's ICE parsing scheme. In S. Greenbaum (Ed.), *Comparing English worldwide: The International Corpus of English* (pp. 125–141). Oxford: Clarendon Press.

Coetzee-Van Rooy, S. (2014). The identity issue in bi- and multilingual repertoires in South Africa: Implications for Schneider's Dynamic Model. In S. Buschfeld, T. Hoffmann, M. Huber, & A. Kautzsch (Eds.), *The evolution of Englishes: The dynamic model and beyond* (pp. 39–57). Amsterdam: John Benjamins.

Davies, M. (2013). *Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries*. Retrieved from https://corpus.byu.edu/glowbe/

Davies, M., & Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-Based English Corpus (GloWbE). *English World-Wide*, *36*(1)1–28.

Deuber, D. (2014). *English in the Caribbean: Variation, style and standards in Jamaica and Trinidad*. Cambridge: Cambridge University Press.

Edwards, A. (2016). *English in the Netherlands: Functions, forms and attitudes*. Amsterdam: John Benjamins.

Edwards, A. (2017). ICE age 3: The Expanding Circle. *World Englishes*, *36*(3), 404–426.

Edwards, A., & Laporte, S. (2015). Outer and Expanding Circle Englishes: The competing roles of norm orientation and proficiency levels. *English World-Wide*, *36*(2), 135–169.

Fang, A. C. (1996). The survey parser: Design and development. In S. Greenbaum (Ed.), *Comparing English worldwide: The International Corpus of English* (pp. 142–160). Oxford: Clarendon Press.

Greenbaum, S. (1988). A proposal for an international computerized corpus of English. *World Englishes*, *7*(3), 315.

Greenbaum, S. (1990a). The International Corpus of English. *ICAME Journal*, *14*, 106–108.

Greenbaum, S. (1990b). Standard English and the International Corpus of English. *World Englishes*, *9*(1), 79–83.

Greenbaum, S. (1990c). The International Corpus of English: A progress report. *World Englishes*, *9*(1), 121–122.

Greenbaum, S. (1991a). ICE: The International Corpus of English. *English Today*, *28*, *7*(4), 3–7.

Greenbaum, S. (1991b). The development of the international corpus of English. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics* (pp. 83–91). London: Longman.

Greenbaum, S. (1992). A new corpus of English: ICE. In J. Svartvik (Ed.), *Directions in corpus linguistics* (pp. 171–179). Berlin: Mouton de Gruyter.

Greenbaum, S. (1993). The tagset for the international corpus of English. In C. Souter & E. Atwell (Eds.), *Corpus-based computational linguistics* (pp. 11–24). Amsterdam: Rodopi.

Greenbaum, S. (1994). Extracts from the annual report (1992–93) of the Survey of English Usage, University College London. *English Today*, *39*, 29–32.

Greenbaum, S. (1996a). Introducing ICE. In S. Greenbaum (Ed.), *Comparing English worldwide: The International Corpus of English* (pp. 3–12). Oxford: Clarendon Press.

Greenbaum, S. (1996b). *Comparing English worldwide: The International Corpus of English*. Oxford: Clarendon Press.

Greenbaum, S., & Ni, Y. (1996). About the ICE tagset. In S. Greenbaum (Ed.), *Comparing English worldwide: The International Corpus of English* (pp. 92–109). Oxford: Clarendon Press.

Gut, U., & Fuchs, R. (2017). Exploring speaker fluency with phonologically annotated ICE corpora. *World Englishes*, *36*(3), 388–403.

Hadikin, G. (2014). *Korean English: A corpus-driven study of a new English*. Amsterdam: John Benjamins.

Hansen, B. (2017). The ICE metadata and the study of Hong Kong English. *World Englishes*, *36*(3), 471–486.

Heller, B., Szmrecsanyi, B., & Grafmiller, J. (2017). Stability and fluidity in syntactic variation world-wide: The genitive alternation across varieties of English. *Journal of English Linguistics*, *45*(1), 3–27.

ICAME. (2010). *ICAME Journal*, *34*. Berlin: Mouton de Gruyter.

Johansson, S., & Stenström, A.-B. (1991). *English computer corpora*. Berlin: Mouton de Gruyter.

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the Outer Circle. In R. Quirk & H. G. Widdowson (Eds.), *English in the world: Teaching and learning the languages and literatures* (pp. 11–30). Cambridge: Cambridge University Press.

Kallen, J. L., & Kirk, J. M. (2008). *ICE-Ireland: A user's guide*. Belfast: Cló Ollscoil na Banríona.

Kallen, J. L., & Kirk, J. M. (2012). *SPICE-Ireland: A user's guide*. Belfast: Cló Ollscoil na Banríona.

Kirk, J. M. (2016). The pragmatic annotation scheme of the SPICE-Ireland corpus. *International Journal of Corpus Linguistics*, *21*(3), 299–322.

Kirk, J. M. (2017). Developments in the spoken component of ICE corpora. *World Englishes*, *36*(3), 371–386.

Kirk, J. M., Kallen, J. L., Lowry, O., Rooney, A., & Mannion, M. (2011). *The SPICE-Ireland corpus: Systems of pragmatic annotation for the spoken component of ICE-Ireland* (Vol. *1.2.2*). VersionBelfast: Queen's University Belfast; Dublin: Trinity College Dublin.

Kirk, J. M., & Nelson, G. (2017). Review of the ICE Project 2016/17. Paper presented in *ICAME38*, Prague, 25 May. [Available from the authors].

Kortmann, B., Burridge, K., Mesthrie, R., Schneider, E. W., & Upton, C. (Eds.). (2004). *A handbook of varieties of English: Morphology & syntax* (Vol. 2). Berlin: Mouton de Gruyter.

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 133–149). Amsterdam: Rodopi.

Lehmann, H.-M. (2015). ICE online. Paper presented at ICAME36, Universität Trier, Germany. 27–31 May.

Lehmann, H.-M., & Schneider, G. (2012). BNC dependency bank 1.0. In S. Oksefjell Ebeling, J. Ebeling, & H. Hasselgård (Eds.), *Aspects of corpus linguistics: Compilation, annotation, analysis* (Vol. 12). Retrieved from http://www.helsinki.fi/varieng/journal/volumes/12/lehmann_schneider/

Leitner, G. (Ed.). (1992a). *New directions in English language corpora: Methodology, results, software developments*. Berlin: Mouton de Gruyter

Leitner, G. (1992b). International corpus of English: Corpus design—problems and suggested solutions. In G. Leitner (Ed.), *New directions in English language corpora: Methodology, results, software developments* (pp. 33–64). Berlin: Mouton de Gruyter.

Loureiro-Porto, L. (2017). ICE vs GloWbE: Big data and corpus compilation. *World Englishes*, *36*(3), 448–470.

Mair, C. (2013). The world system of Englishes: Accounting for the transnational importance of mobile and mediated vernaculars. *English World-Wide*, *34*(3), 253–278.

Mair, C. (2015). Response to Davies and Fuchs. *English World-Wide*, *36*(1), 29–33.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Mukherjee, J. (2015). Response to Davies and Fuchs. *English World-Wide*, *36*(1), 34–37.

Nelson, G. (1991). Markup for spoken texts. *ICE newsletter 10*. London: Survey of English Usage.

Nelson, G. (1995). The International corpus of English: Markup for spoken language. In G. Leech, G. Myers, & J. Thomas (Eds.), *Spoken English on computer: Transcription, mark-up and application* (pp. 220–223). London: Longman.

Nelson, G. (1996). Markup systems. In S. Greenbaum (Ed.), *Comparing English worldwide: The International Corpus of English* (pp. 36–53). Oxford: Clarendon Press.

Nelson, G. (2002a). International corpus of English markup manual for spoken texts. Retrieved from http://www.ice-corpora.uzh.ch/en.html

Nelson, G. (2002b). International corpus of English markup manual for written texts. Retrieved from http://www.ice-corpora.uzh.ch/en.html

Nelson, G. (2015). Response to Davies and Fuchs. *English World-Wide*, *36*(1), 34–40.

Nelson, G. (2017). The ICE project and world Englishes. *World Englishes*, *36*(3), 367–370.

Nelson, G., Wallis, S., & Aarts, B. (2002). *Exploring natural language: Working with the British component of the International Corpus of English*. Amsterdam: John Benjamins.

Oostdijk, N. (1991). *Corpus linguistics and the automatic analysis of English*. Amsterdam: Rodopi.

Ozón, G., Ayafor, M., Green, M., & Fitzgerald, S. (2017). A spoken corpus of Cameroon Pidgin English. *World Englishes*, *36*(3), 427–447.

Peters, P. (2015). Response to Davies and Fuchs. *English World-Wide*, *36*(1), 41–44.

Porter, N., & Quinn, A. (1996). Developing the ICE corpus utility programme. In S. Greenbaum (Ed.), *Comparing English worldwide: The International Corpus of English* (pp. 79–91). Oxford: Clarendon Press.

Quinn, A., & Porter, N. (1996). ICE annotation tools. In S. Greenbaum (Ed.), *Comparing English worldwide: The International Corpus of English* (pp. 65–78). Oxford: Clarendon Press.

Rayson, P., Piao, S., & Archer, D. (2004). The UCREL semantic analysis system. In Proceedings of the workshop on *Beyond Named Entity Recognition Semantic labelling for NLP tasks in association* with *4th International Conference on Language Resources and Evaluation* (LREC 2004), Lisbon, Portugal, 25 May, 2004, 7–12. http://ucrel.lancs.ac.uk/usas

Saraceni, M. (2015). *World Englishes: A critical analysis*. London: Bloomsbury.

Schmied, J. (1996). Second language corpora. In Greenbaum (Ed.), *Comparing English worldwide: The International Corpus of English* (pp. 182–196). Oxford: Clarendon Press.

Schneider, E. (2007). *Post-colonial Englishes: Varieties around the world*. Cambridge: Cambridge University Press.

Schneider, G. (2008). *Hybrid long-distance functional dependency parsing* (PhD thesis). University of Zurich, Zurich. Retrieved from https://files.ifi.uzh.ch/cl/CLpublications/schneider-diss.pdf

Searle, J. R. (1969). *Speech acts*. Cambridge: Cambridge University Press.

Searle, J. R. (1976). A classification of illocutionary speech acts. *Language in Society*, *5*(1), 1–23.

Spolsky, B. (2003). *Language policy*. Cambridge: Cambridge University Press.

Szmrecsanyi, B., Grafmiller, J., Heller, B., & Röthlisberger, M. (2016). Around the world in three alternations: Modeling syntactic variation in varieties of English. *English World-Wide*, *37*(2), 109–137.

Wallis, S. (n.d.). Tagging ICE-Philippines and other Corpora. Retrieved from http://www.ucl.ac.uk/english-usage/staff/sean/resources/tagging-ice-phi.pdf

Wallis, S., Aarts, B., & Nelson, G. (2000). Parsing in reverse: Exploring ICE-GB with fuzzy tree fragments and ICE-CUP. In J. M. Kirk (Ed.), *Corpora galore: Analyses and techniques in describing English* (pp. 335–344). Amsterdam: Rodopi.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation* (LREC 2006) (pp. 1556–1559). Retrieved from http://www.lrec-conf.org/lrec2006

Wong, D., Cassidy, S., & Peters, P. (2011). Updating the ICE annotation system: Tagging, parsing and validation. *Corpora*, *6*(2), 115–144.

World Englishes. (1996). Special issue on ICE project. *World Englishes*, *15*(1). Oxford: Wiley.

World Englishes. (2004). Special issue on ICE project. *World Englishes*, *23*(2). Oxford: Wiley.

World Englishes. (2017). Special issue on ICE project. *World Englishes*, *36*(3). Oxford: Wiley.

Wunder, E.-M., Voormann, H., & Gut, U. (2010). The ICE Nigeria corpus project: Creating an open, rich and accurate corpus. *ICAME Journal*, *34*, 78–88.

Zeldes, A., Ritz, J., Lüdeling, A., & Chiarcos, C. (2009). ANNIS: A search tool for multi-layer annotated corpora. In M. Mahlberg, V. González-Díaz, & C. Smith (Eds.), *Proceedings of corpus linguistics* (Vol. 2009, Article 358). Retrieved from http://ucrel.lancs.ac.uk/publications/cl2009/

How to cite this article: Kirk J, Nelson G. The International Corpus of English project: A progress report. *World Englishes*. 2018;37:697–716. https://doi.org/10.1111/weng.12350

APPENDIX 1

TABLE A1 Text categories in ICE

Spoken text categories			N.
Dialogue (180)	Private (100)	Direct conversations	90
		Distanced (telephone) conversations	10
	Public (80)	Class lessons	20
		Broadcast discussions	20
		Broadcast interviews	10
		Parliamentary debates	10
		Legal cross-examinations	10
		Business transactions	10
Monologue (120)	Unscripted (70)	Spontaneous commentaries	20
		Unscripted speeches	30
		Demonstrations	10
		Legal presentations	10
	Scripted (50)	Broadcast news	20
		Broadcast talks	20
		Speeches (not broadcast)	10
Total			300

Written text categories			N.
Non-printed (50)	Non-professional writing (20)	Student untimed essays	10
		Student examination essays	10
	Correspondences (30)	Social letters	15
		Business letters	15
Printed (150)	Informational (learned) (40)	Humanities	10
		Social Sciences	10
		Natural Sciences	10
		Technology	10
	Informational (popular) (40)	Humanities	10
		Social Sciences	10
		Natural Sciences	10
		Technology	10
	Informational (reportage) (20)	Press news reports	20
	Instructional (20)	Administrative/regulatory prose	10
		skills/hobbies	10
	Persuasive (10)	Press editorials	10
	Creative (20)	Novels/short stories	20
Total			200

APPENDIX 2: ICE-CORPORA & TEAMS: updated to August 2017

* indicates that the component is included in ICE Online at the University of Zurich (ICE9 and ICE15)

Teams declared inactive following release of corpus

1. ICE-India*
   Prof. Gerhard Leitner, leitner@philologie.fu-berlin.de (Freie Universität Berlin)
2. ICE-New Zealand*
   Dr Bernadette Vine, bernadette.vine@vuw.ac.nz (Victoria University of Wellington)
3. ICE-Singapore*
   Dr Vincent Ooi, vinceooi@nus.edu.sg (National University of Singapore)

Active Teams, corpus compiled, released and ‘still maintained’

1. ICE-Australia*
   Prof. Pam Peters, pam.peters@mq.edu.au (Macquarie University, Sydney)
2. ICE-Canada*
   Professor John Newman, john.newman@ualberta.ca (University of Alberta)
3. ICE-GB*
   Prof. Bas Aarts (UCL), b.aarts@ucl.ac.uk and Dr Sean Wallis (UCL) s.wallis@ucl.ac.uk
4. ICE-East Africa (Kenya and Tanzania)
   Prof. Josef Schmied, josef.schmied@phil.tu-chemnitz.de (Technische-Universität Chemnitz)
5. ICE-Hong Kong*
   Prof. Kingsley Bolton, kbolton@ntu.edu.sg (Nanyang Technological University Singapore)
6. ICE-Ireland* (also SPICE-Ireland)*
   Prof. Dr John Kirk, jk@etinu.com (formerly Queen's University Belfast, now University of Vienna) and Dr Jeffrey Kallen, john.kirk@univie.ac.at (Trinity College Dublin)

7. ICE-Jamaica*
   Prof. Dr Christian Mair (Universität Freiburg), mairch@ruf.uni-freiburg.de and Dr K. Shields-Brodber and Prof. Dr Hubert Devonish (University of the West Indies, Kingston)
8. ICE-Nigeria*
   Prof. Dr Ulrike Gut, gut@uni-muenster.de (Westfälische Wilhemsuniversität Münster), Prof. Dr Inyang Udofot (University of Uyo, Akwa Ibom State) and Prof. David Jowitt (University of Jos)
9. ICE-Philippines*
   Dr Ariane Borlongan, arianemacalingaborlongan@yahoo.com (University of Tokyo)

Written components released

1. ICE-Ghana* (spoken to follow 2018)
   Prof. Dr Magnus Huber, magnus.huber@anglistik.uni-giessen.de (Universität Giessen) and Prof. Kari Dako (University of Ghana)
2. ICE-Sri Lanka* (spoken to follow 2018)
   Prof. Dr Joybrato Mukherjee, mukherjee@uni-giessen.de and Dr Tobias Bernaisch (Universität Giessen) Tobias.J.Bernaisch@anglistik.uni-giessen.de
3. ICE-USA* (no spoken component)
   Prof. Charles F. Meyer, meyer@cs.umb.edu (University of Massachusetts-Boston)

ICE- Corpora in compilation

4. ICE-Bahamas (no date)
   Prof. Dr Stephanie Hackert, stephanie.hackert@anglistik.uni-muenchen.de (Ludwig-Maximilian-Universität, Munich)
5. ICE-Fiji* (no date)
   Prof. Dr Marianne Hundt, m.hundt@es.uzh.ch (Universität Zurich) and Prof. Dr Carolin Biewer carolin.biewer@uni-wuerzburg.de (Universität Würzburg) and Dr Jan Tent (Macquarie University, Sydney)
6. ICE-Gibraltar (not before 2020)
   Dr Elena Seoane, elena.seoane@uvigo.es (University of Vigo), Dr Cristina Suárez-Gómez, cristina.suarez@uib.es and Lucía Loureiro Porto, lucia.loureiro@uib.es (University of the Balearic Islands)
7. ICE-Malaysia (by 2020)
   Dr Hajar Abdul Rahim, hajar@usm.my (Universiti Sains Malaysia, Penang) and Dr Su'ad Awab (The University of Malaya, Kuala Lumpur)
8. ICE-Malta (no date)
   Prof. Manfred Krug, manfred.krug@split.uni-bamberg.de and Dr Ole Schützler, ole.schuetzler@split.uni-bamberg.de (Universität Bamberg)
9. ICE-Puerto Rico (no date)
   Prof. Manfred Krug, manfred.krug@uni-bamberg.de (Universität Bamberg) and Prof. Dr Don E. Walicek (University of Puerto Rico, San Juan)
10. ICE-Scotland (2017–2018)
    Prof. Dr Ulrike Gut, gut@uni-muenster.de (Westfälische Wilhemsuniversität Münster), Prof. Dr Robert Fuchs, robert.fuchs@uni-hamburg.de (Universität Hamburg) and Dr Ole Schützler, ole.schuetzler@split.uni-bamberg.de (Universität Bamberg)
11. ICE-South Africa (written texts by late 2017)
    Prof. Dr Bertus van Rooy, Bertus.vanRooy@nwu.ac.za (North Western University)

12. ICE Trinidad & Tobago (2018)
    Prof. Dr Dagmar Deuber, deuber@uni-muenster.de (Westfälische Wilhemsuniversität Münster) and Prof. Valerie Youssef (University of the West Indies)
13. ICE-Uganda (written 2017; spoken no date)
    Prof. Dr Christiane Meierkord, christiane.meierkord@rub.de (Ruhr-University of Bochum)

ICE-corpora the status of which is uncertain, as no responses to the questionnaire were received

1. ICE-Namibia (no date, no response)
   Prof. Sarala Krishnamurthy, skrishnamurthy@polytechnic.edu.na (Polytechnic of Namibia, Windhoek)
2. ICE-Pakistan (no date, no response)
   Rashid Mahmood, ch.raashid@gmail.com and Asim Mahmood (University Faisalabad)

APPENDIX 3: ELECTRONIC TEXTS

The review recommends the inclusion of a third textual component in a second generation ICE corpus: electronic texts. However, we did not feel that it was a remit of our review to set down exactly which electronic text categories or in what amounts they should be included. However, we offer the following document as a basis for further discussion.

A distinction has to be made between electronic text genres that originate in speech or writing, and which have come to have electronic forms (for example newspaper reports or editorials in writing; demonstrations in speech), and electronic text genres that originated electronically, and only exist electronically (for example Twitter). The former should remain as written texts and be collected alongside the written equivalents; while only the latter would be collected within the component 'electronic texts'. As one respondent urged: 'It would be better to include electronic forms of the old functions in the old text categories even if the total number of texts in such a category were expanded'. It is of course entirely legitimate for written texts to be downloaded from the Internet. A newspaper report, for instance, remains a written text, even if collected in this way. However, what may be important in preparing this text for inclusion in a corpus is that any advertising or boilerplate material present on the page, as well as features related to HTML codes, such as hyperlinked URLs, are removed very carefully. Although proposals for taxonomies of electronic texts/social media data/computer-mediated communication have been made (e.g. Biber & Kurjian 2007), what follows is a proposal of our own. Electronic texts may be subdivided into four categories: extended-written; written-like, spoken-like, and multi-media.

A3.1 Extended written texts

Extended written texts are the written texts of *websites* of various kinds, often also containing multi-media: (public) institutional/administrative websites; corporate sites; commercial sites; cultural websites; clubs and societies websites; and so on. These are mostly equivalent to monologic, informational texts.

A3.2 Written-like in mode form and function, unlikely to have been professionally edited: the electronic medium begat the text category

There could be two sub-categories: *written-formal* and *written-informal*. *Written-formal* approximates to formal uses of writing (for example informational, transactional) as well as to norms of writing in such contexts. *Written-informal* approximates to social communication for purpose of maintaining good social relationships, using colloquial, oral and other linguistic devices marking informality.

E-mails are asynchronous written messages between people, sent to known or identifiable recipients using digital devices such as computers, tablets and mobile phones. They are generally similar in nature to many formal types of written communication (such as letters and memos), but, as they are also used for informal communication, may not always follow the same rules of formality in terms of capitalization, spelling, and so on, and may thus also have a more 'oral' character. In terms of corpus processing, it may be necessary to identify and remove quoted content repeated from the original message. However, one problem with this may be that a writer may use such quoted materials in an attempt to create 'synchronicity' with the original sender by referring to all or parts of the original message verbatim, rather than referring to the content by paraphrasing what is being responded to. *Texts or SMSs (Short Message Service)* are asynchronous written messages using mobile telephony systems. The activity is often referred to as 'texting'. Due to limited space and potential cost, these texts exhibit cost-saving features, such as employing 'telegraph style', acronyms or other types of 'abbreviations' (for example LoL, *4* instead of *for*, *Ur* for *your*, *and btw* for *'by the way'*). *Tweets are* asynchronous written messages using the social networking service Twitter, where users post and interact with written messages, each tweet message being restricted to 140 characters. Similar to above, for similar reasons.

A3.3 Spoken-like in mode form and function – monologic and dialogic text categories extended to the electronic medium

Chat may refer to short written messages conveyed over the Internet on a dialogic basis between a sender and receiver. Chat messages are generally short in order to enable other participants to respond quickly, thereby creating a feeling similar to a spoken conversation. Although chats are designed to allow synchronous communication, asynchronicity may be often be introduced, due to one 'interlocutor' typing faster than another, with the other not responding quickly enough. In other words, even though chats are designed to produce 'adjacency pairs', there is often no such regularity between initiation and response. Colloquial, informal features of spontaneous discourse abound, as well as acronyms and other cost-saving devices of the electronic medium. *Discussion groups* are used by individuals to exchange written comments in an interactive, dialogic, but asynchronous, way. They are often separated into individual threads, so that topics generally remain consistent. As contributions may contain considered responses, more typical of written language, they may have correspondingly fewer colloquial features. Regarding corpus processing, though, a similar issue with repeated/quoted content may exists.

A3.4 Multi-media—because multi-media, the electronic medium begat the text categories —crucially linkage of websites and video with text or speech

Facebook postings are asynchronously exchanged written and often visual messages between people online using this particular social media/social networking service. *Skype* (*skypeing*) is an online application which enables direct dialogic spoken, written (chat) and video exchanges of any length, and also includes options for leaving voice messages. *Blogs* (< *weblogs*) are written messages typically like a narrative to inform about personal information or a commentary or position-statement to articulate views about a topic of current public interest, usually written in an informal, colloquial style, and issued relatively frequently, sometimes daily, and usually dated. Blogs may include images or illustrations or embed video material. Some blogs enable readers to respond or comment. Blogs are often thought of as the most frequent or typical use of Internet communication. *Vlogs* (< *video blog*) and *podcasts* (a portmanteau of *ipod* and *broadcast*) are spoken forms of blogs and may embed video or have supporting text, images, and other metadata. *Citizen broadcasting* are spoken video broadcasts transmitted by individuals across the internet. They maybe be monologic or involve interaction between the participant speakers. They are not addressed to a particular audience. *Text Length*—Electronic texts are relatively short, so that many 2,000 word 'texts' will be composite texts. There are many views about length of individual texts with regard to balance and representativeness but we feel that the 2,000 word sample—whether only an excerpt in the case of a much longer work or in collations of (say) text messages (SMSs) or tweeks—remains useful as a standard, comparable unit (certainly over 'whole' texts, regardless of length, as advocated by some). See Table A2.

TABLE A2 Proposal: given the above, we propose that the following types and quantities of electronic text be collected in second generation corpora

Electronic Texts	No of texts × length	No. of words
A3.1 Extended written texts	30	60,000
A3.2 Written-like	30	60,000
A3.3 Spoken-like	80	160,000
A3.4 Multi-media	110	220,000
Total	250	500,000

In more detail:

Electronic Texts	No of texts × length	No. of words
Extended written texts	30	60,000
Websites	30 × 2,000	
Written-like	30	60,000
E-mails	10 × 2,000	
Texts or SMSs	10 × 2,000	
Tweets	10 × 2,000	
Spoken-like	80	160,000
Chat	40 × 2,000	
Discussion groups	40 × 2,000	
Multi-media	110	220,000
Facebook postings	20 × 2,000	
Skype (skypeing)	20 × 2,000	
Blogs	40 × 2,000	
Vlogs and podcasts	20 × 2,000	
Citizen broadcasting	10 × 2,000	
Total	250	500,000

Having said that, if we are applying the *flexibility principle* to the inclusion of text categories and to the number of texts in a category, then it follows that the principle should apply to text length. Instead of creating composite texts, for each electronic text category, the total number of words (say 20,000 words) could simply comprise all the texts individually which make up that total—if an SMS contains an average of 20 words, it would simply be a question of collecting 1000 SMSs individually, without recourse to ad hoc composite groupings of c. 2,000 words per grouping.