41 World Englishes and Corpora Studies

GERALD NELSON

1 Introduction

Kennedy (1998: 1) provides succinct definitions of both the terms "corpus" and "corpus linguistics":

> In the language sciences a corpus is a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description. Over the last three decades the compilation and analysis of corpora stored in computerized databases has led to a new scholarly enterprise known as corpus linguistics.

As Kennedy's definition shows, the corpus-based method of linguistic research is a very recent development, and the use of corpora in the study of world Englishes is more recent still. McEnery and Wilson (2001: 1) provide a useful account of what they call "early" corpus linguistics, by which they refer to range of research projects undertaken from the 1950s to the 1970s, using entirely manual methods for compiling and analyzing large collections of text. Notable among these was the work of Randolph Quirk, who compiled the Survey of English Usage (SEU) corpus, beginning in 1959. SEU is a one-million-word corpus of British English, dating for the most part from the 1960s. From our perspective in the technologically sophisticated twenty-first century, it is astonishing to recall that the SEU corpus was an entirely paper-based corpus, with each instance of every word annotated on its own paper slip, and the slips stored in a vast array of metal filing cabinets (Peppé, 1995). In contrast with this, corpus linguistics today exploits the ever-increasing power of computer hardware and software, and with the aid of computer technology, linguists are compiling ever-larger collections of text. Since the 1980s, the corpus-based approach has become firmly established as a methodology for linguistic research.

Table 41.1 Composition of the Brown Corpus

Informative prose: 374 samples	Imaginative prose: 126 samples
Press: reportage	General fiction
Press: editorial	Mystery and detective fiction
Press: reviews	Science fiction
Religion	Adventure and western fiction
Skills and hobbies	Romance and love story
Popular lore	Humor
Belles lettres, biography, memoir	
Miscellaneous	
Learned	

2 Electronic Corpora

The first electronic corpus of English is generally agreed to be the Brown corpus, which was compiled by Francis and Kučera at Brown University, Rhode Island, in 1963–4. The compilers refer to the corpus as *A Standard Corpus of Present-Day Edited American English* (Francis and Kučera, 1971). It consists of just over one million words of printed English produced in the United States during the calendar year 1961. It includes 500 individual samples of 2,000 words each, selected from the range of text types shown in Table 41.1.

The Brown corpus has been, and continues to be, enormously influential, especially in terms of the methodology of corpus design and compilation. For that reason, it is worth quoting the compilers at some length here:

> Samples were chosen for their representative quality rather than for any subjectively determined excellence. The use of the word standard in the title of the Corpus does not in any way mean that it is put forward as "standard English"; it merely expresses the hope that this corpus will be used for comparative studies where it is important to use the same body of data. Since the preparation and input of data is a major bottleneck in computer work, the intent was to make available a carefully chosen and prepared body of material of considerable size in standardized format. The corpus may further prove to be standard in setting the pattern for the preparation and presentation of further bodies of data in English or in other languages. (Francis and Kučera, 1971)

The Brown corpus did, indeed, become a "standard" in the sense that the compilers express here. It set in motion a series of corpus-based projects around the world, in which the researchers invariably looked to Brown as their model. The Lancaster-Oslo/Bergen (LOB) corpus was begun in 1976 in order to provide a British English equivalent of the Brown corpus (Johansson, Leech, and Goodluck, 1978). To this end, the compilers followed the design of Brown closely, selecting texts printed in Great Britain in 1961, and choosing the same number and size of samples from the same text categories. The objective was, of course, to ensure that the two corpora would be directly comparable with each other, so that they could be used as the basis for comparative studies across the two dominant varieties, American and British English (AmE and BrE).

In 1978, S. V. Shastri noted that previous studies of Indian English had been largely confined to aspects of the spoken variety (Bansal, 1969), or to isolated topics in the language (Kachru, 1965). Having worked at Lancaster University with Geoffrey Leech, one of the prime movers behind the LOB corpus, Shastri recognized that "a comprehensive description [of Indian English] will have to be based on a standard corpus" (Shastri, 1986). To this end, Shastri compiled the Kolhapur corpus of written Indian English, using both Brown and LOB as his models. He declared his objectives in the following terms:

> The present corpus of Indian Written English is comparable to the Brown and the LOB corpora. It is intended to serve as source material for comparative studies of American, British and Indian English which in its turn is expected to lead to a comprehensive description of Indian English. (Shastri, 1986)

However, unlike Brown and LOB, which sampled texts from 1961, the Kolhapur corpus takes 1978 as its sampling date. Part of the rationale behind this had to do with the perceived "Indianness" of post-independence Indian English. As Shastri explained:

> it is felt that the value of the Indian corpus is immensely enhanced in general and in particular as a source for the description of Indian English . . . as the Indianness of Indian English is a post-independence phenomenon and may have reached a discernible stage in the thirty years after Independence. It is argued in theory that in the same thirty years the American and British English may not have undergone such changes. (Shastri, 1986)

This is an interesting observation, and one which, consciously or unconsciously, informs descriptions of other post-colonial Englishes as well. What Shastri was consciously attempting to construct was a corpus of distinctively Indian English, as opposed to the variety used at the time of Independence. Whether the 30-year gap to 1978 would be sufficient to allow the "Indianness" of Indian English to manifest itself is perhaps a moot point. The key issue here is that Shastri, following Kachru (1965), recognized Indian English as a distinct variety, and set about capturing it in the Kolhapur corpus.

The Australian Corpus of English (ACE) was compiled at Macquarie University, beginning in 1986. As with the Kolhapur corpus, the compilers were motivated primarily by a wish to differentiate between their own variety of English and the British and American varieties. For that reason, they followed the Brown and LOB models closely in terms of corpus design, though again there is a chronological gap: ACE samples texts from 1986.

At Victoria University of Wellington, New Zealand, researchers compiled the Wellington Corpus of Written New Zealand English (WWC; Bauer, 1993). Once again, Brown and LOB were the models, though the compilers decided to use ACE as their model in terms of sampling date. In the Wellington corpus, the majority of samples date from 1986 or 1987. The corpus of written New Zealand English was followed in 1998 by the Wellington Corpus of Spoken New Zealand English (WSC), consisting of dialogs and monologs collected in the period 1988 to 1994 (Holmes, Vine, and Johnson, 1998).

Beginning with the highly influential Brown corpus in the early 1960s, the enterprise of compiling English-language corpora has continued in highly principled and systematic ways. As a result, linguists now have five "parallel" corpora of international written English at their disposal: Brown, LOB, Kolhapur, ACE, and Wellington. In 1990, a new project was initiated which would significantly expand this collection, and more importantly, greatly increase both the linguistic and the geographical coverage of available corpora.

3 The International Corpus of English

The International Corpus of English (ICE) project was conceived in the late 1980s by Sidney Greenbaum, then Director of the Survey of English Usage, University College London. The idea was first proposed in a brief notice in *World Englishes* (Greenbaum, 1988), in which researchers were invited to collaborate on the compilation of parallel English corpora, specifically in countries where English is used as a first language, or as a second official language. The invitation was timely, and the response from linguists worldwide was both immediate and enthusiastic. The ICE project currently involves research teams working in the following countries or regions: Australia, Canada, East Africa (Kenya and Tanzania), Great Britain, Hong Kong, India, Ireland, Jamaica, Malaysia, New Zealand, Philippines, Singapore, South Africa, Sri Lanka, United States.

From its inception, ICE aimed to compile parallel corpora from two of Kachru's Three Circles of English (Kachru, 1985). The Inner Circle is represented by countries such as Britain, the United States, and Australia, while the Outer Circle is represented by countries such as India, Singapore, and the Philippines. Kachru's third circle, the Expanding Circle, is represented in an anciliary project, the International Corpus of Learner English (ICLE), which is discussed below.

Each ICE team is compiling (or has already compiled) a one-million-word corpus of their own variety of English, produced by adults (aged 18 or over) in the period after 1989. While each national or regional corpus can exist independently as a valuable resource for the study of individual varieties, the real value of the corpora lies in their being exactly compatible with each other. This compatibility lies in every area of the corpus design and annotation (Nelson, 1996a, 1996b). The design, in terms of text categories, is shown in Table 41.2.

Each corpus consists of 500 samples of approximately 2,000 words each, to give a total of one million words. The first major division is between speech

Table 41.2 Composition of the ICE corpora

WRITTEN TEXTS (200 samples)	SPOKEN TEXTS (300 SAMPLES)
Non-printed	**Dialogue**
Non-professional writing	*Private*
Student essays	Direct conversations
Examination scripts	Telephone calls
Correspondence	*Public*
Social letters	Class lessons
Business letters	Broadcast discussions
Printed	Broadcast interviews
Academic writing	Parliamentary debates
Humanities	Legal cross-examinations
Social sciences	Business transactions
Natural sciences	**Monologue**
Technology	*Unscripted*
Non-academic writing	Spontaneous commentaries
Humanities	Unscripted speeches
Social sciences	Demonstrations
Natural sciences	Legal presentations
Technology	*Scripted*
Reportage	Broadcast news
Press news reports	Broadcast talks
Instructional writing	Non-broadcast talks
Administrative writing	
Skills and hobbies	
Persuasive writing	
Press editorials	
Creative writing	
Novels and stories	

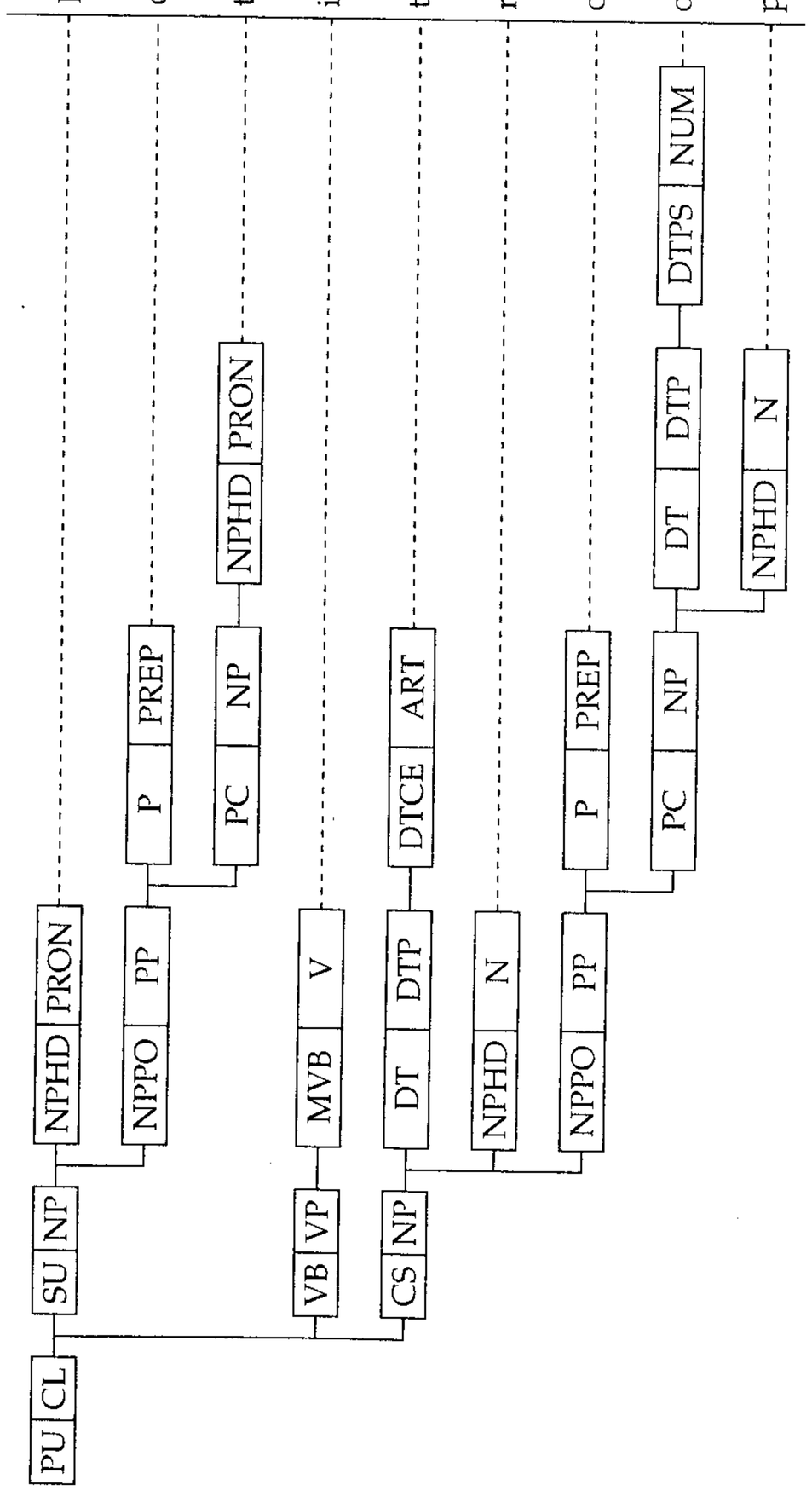
(300 samples) and writing (200 samples). Further subdivisions are made in a hierarchical fashion, with speech divided among dialog (180 samples) and monolog (120 samples), and writing divided among non-printed (50 samples) and printed (150 samples). The hierarchical subdivision continues to the fundamental level of the text categories, of which there are 15 in speech and 17 in writing.

The overall design was arrived at following extensive discussion (see Leitner, 1992; Schmied, 1990). While it is informally based on the design of Brown and LOB, it also reflects some important differences. Most notably, it samples spoken English, and in a greater proportion than writing. Within the spoken component,

by far the greatest contribution is from face-to-face conversations (90 samples, or 180,000 words). The ICE corpora, therefore, are distinctive in the emphasis they place on the spoken medium, and in particular on informal, conversational English. They are also distinctive in that they include only those text categories which are internationally applicable. So, for example, the corpus design (Table 41.2) contains no Religion category, as Brown and LOB did, because writing on this topic is not available (in English, at least) in all the participating countries. Similarly, the subdivision of Fiction into Romance, Westerns, Detective Fiction, etc., have been dispensed with, since these subtypes simply do not apply internationally. The ICE corpora aim to be maximally representative of English in use in all the participating countries, and not in any one country.

Each of the ICE teams has had a slightly different experience in compiling their respective corpora, depending on local circumstances, and specifically on the status of English in the country concerned. Many of the teams have written informatively about these experiences, and they provide some valuable insights into the processes and rationales behind corpus building in the context of world Englishes. Schmied (1995) discusses the issue of national standards in the context of the ICE project, with special emphasis on East African varieties. Holmes (1996) discusses methodological problems in compiling the spoken component of ICE New Zealand. Bolt and Bolton (1996) discuss the Hong Kong ICE project, and their observations provide an interesting contrast with Shastri's comments on the "Indianness" of Indian English, discussed earlier. The Hong Kong component of ICE (ICE-HK) was compiled during a crucial period in the territory's history: the texts in the corpus date from both before and after the "Handover" in 1997, which saw Hong Kong reverting from British to Chinese rule. As such, the data in the corpus may be said to represent English in Hong Kong at the "end of empire," in contrast with Shastri's conception of Indian English 30 years after Independence. Indeed, the compilers of ICE-HK worked in the knowledge that this would probably be the last opportunity to sample "Hong Kong English," since they predicted (correctly, it now seems) that the status of English in Hong Kong was about to change dramatically, with both Cantonese and Putonghua (Mandarin) rising quickly to prominence (Bolt and Bolton, 1996). Thus, while conforming to an agreed international standard, each ICE corpus reflects the unique situation of English in each participating region.

At the time of writing, six ICE corpora are available for the purposes of non-profit, academic research. These are Great Britain (ICE-GB), New Zealand (ICE-NZ), Singapore (ICE-SIN), India (ICE-IND), East Africa (ICE-EA), and the Philippines (ICE-PHI). Details of availability are given at the end of this chapter. The British corpus, ICE-GB, was completed in 1998. In terms of annotation, it is the most advanced of all the ICE corpora. Every word has been tagged for part of speech, using a specially designed tagset (Greenbaum, 1993), and each sentence/utterance has been parsed at phrase and clause level. The syntactic structures are represented in the familiar form of tree diagrams, as illustrated in Figure 41.1.



Key: ART: article; CL: clause; CS: subject complement; DT: determiner; DTCE: central determiner; DTP: determiner phrase; DTPS: post-determiner; N: noun; NP: noun phrase; NPHD: noun phrase head; NPPO: noun phrase postmodifier; MVB: main verb; NUM: numeral; P: prepositional; PC: prepositional complement; PP: prepositional phrase; PREP: preposition; PRON: pronoun; PU: parsing unit; V: verb

Figure 41.1 A tree diagram from the ICE-GB corpus

The ICE-GB corpus contains just over 83,000 syntactic trees, and represents the largest amount of data ever parsed to this level of detail. The grammatical terminology used in the corpus is based for the most part on the function/form approach found in *A Comprehensive Grammar of the English Language* (Quirk et al., 1985). The annotation was carried out using software developed by the TOSCA research group at the University of Nijmegen, under the direction of Professor Jan Aarts (Van Halteren and Oostdijk, 1993). The corpus is distributed with its own retrieval software, ICECUP (the ICE Corpus Utility Program), which supports complex searches of the syntactic trees (Nelson, Wallis, and Aarts, 2002).

One of the central principles underlying the ICE project is the notion that a common grammatical "core" unites all the varieties. This can be seen, perhaps, as a slight but important shift of emphasis: from concentrating on the "distinctiveness" of, say, Indian English or Australian English, ICE focuses primarily on what unites the varieties. The notion of the common core is described in the following terms by Quirk et al. (1985: 16):

> A common core or nucleus is present in all the varieties so that, however esoteric a variety may be, it has running through it a set of grammatical and other characteristics that are present in all the others. It is this fact that justifies the application of the name "English" to all the varieties.

Quirk et al. appear to accept the existence of the common core as an established fact, though this cannot be empirically tested without extensive study of parallel corpora such as the ICE components. Discovering whether such a core actually exists or not, and how it might be constituted, is perhaps the ultimate objective of the ICE project. In order to do this, all the ICE corpora will have to be annotated to the same level as ICE-GB. That is, they will all have to be syntactically parsed at least to the level of annotation illustrated in Figure 41.1. Once this has been achieved, linguists will be able to examine those grammatical structures that form the putative "core," and also to see which structures (if any) are present only in individual varieties. However, the other currently available corpora mentioned above are still at the "lexical" stage; that is, they contain the words only, with no part of speech tagging or syntactic analysis. Reaching the level of annotation of ICE-GB will be time-consuming and expensive, but well worth the effort. However, despite their lack of annotation at present, many of the ICE corpora have already proved invaluable as sources of data for comparative studies of English varieties.

4 Corpus-Based Studies of World Englishes

The corpora described above have been used as the basis for a very large and varied body of research, and they continue to be used in this way. The results of this research are far too numerous to cite here, though special mention

should perhaps be made of the pioneering work, *Computational Analysis of Present-Day American English* (Kučera and Francis, 1967), based on the Brown corpus. This large-scale, quantitative study was later replicated using the LOB corpus of British English, in *Frequency Analysis of English Vocabulary and Grammar* (Johansson and Hofland, 1989). For comprehensive bibliographies of corpus-based studies of English, see Glauser, Schneider, and Görlach (1993), Altenberg (1998), and Fallon (2004).

The Kolhapur corpus has proved to be an especially fruitful resource for investigators of world Englishes. For an account of early work based on the corpus, see Shastri (1988). Sayder (1989) contrasts the use of the subjunctive in Indian, British, and American English, while Leitner (1992) analyzes the verbs *begin* and *start* in Indian English, in comparison with both AmE and BrE. A particularly important recent contribution in this context is Schneider (2000), which analyzes a range of grammatical phenomena in the Kolhapur corpus, including the subjunctive, case marking of *wh*-pronouns, pro-form *do*, and the indefinite pronouns in *-body* and *-one*. Comparing his findings with those from Brown and LOB, Schneider concludes that "my empirical corpus investigations have shown that no fundamental, categorical difference between Indian English and any other of the national varieties was detected, but on the other hand there is no full identity of patterns and preferences to be observed" (Schneider, 2000: 133). Schneider's conclusion is a complex one, and in particular his finding that "no full identity of patterns" may be observed is especially interesting in the light of the putative common core.

The ACE corpus of Australian English has provided data for a wide range of investigations, focusing on, for example, comparisons of Australian and British usage (Peters 1993a, 1993b), the influence of AmE and BrE on Australian verb morphology (Peters, 1994), the language of Australian newspapers (Peters et al., 1988), and the semantics of modal verbs (Collins, 1988, 1991). The two Wellington corpora – of written New Zealand English (WWC) and of spoken New Zealand English (WSC) – have supported research into a wide range of topics, including gender-based variation (Holmes, 1993), relative pronouns (Sigley, 1997), and the discourse of direct and indirect speech (Yang, 1997).

Whether researchers examine aspects of Indian English, Australian English, or New Zealand English, a comparison is usually made – explicitly or implicitly – with AmE and/or BrE. This is to be expected, because of the traditional dominance of these two varieties, and, on a more practical level, because of the availability of the Brown and LOB corpora. The corpora in the International Corpus of English offer scope for much more inclusive studies of English worldwide, taking account not only of first-language varieties, but of second-language varieties as well. ICE-GB has been extensively used in research, initially as a "snapshot" of BrE in the early 1990s, and later in comparative studies with other varieties. Most notably, ICE-GB formed the most important data source for the *Oxford English Grammar* (Greenbaum, 1996), and a subset of the corpus was used in a study of subordination in speech and

writing (Greenbaum and Nelson, 1995a, 1995b, 1996; Greenbaum, Nelson, and Weitzman, 1995).

A range of "first findings" from other ICE corpora were published in a special issue of *World Englishes* in 1996 (vol. 15, no. 1). The papers in that volume deal with such topics as coordination in BrE and AmE (Meyer, 1996), the language of sports reporting (Leitner and Hesselmann, 1996), intervocalic /t/ in New Zealand English (Bauer and Holmes, 1996), and the stylistic features of East African newspaper English (Schmied and Hudson-Ettle, 1996). Since then, the task of completing the ICE corpora has continued, and researchers have continued to explore the data as it has been collected. Sand (1998) examined the structure of the verb phrase in Jamaican English, and found that while it does not deviate markedly from international Standard English norms, the preferences selected by users in given situations are markedly different. For instance, in the expression of future time, Sand found that both the *will* form (including *'ll* and *shall*) and the *going to* form (including *gonna*) are readily attested in Jamaican speech, just as they are in BrE. However, she found that the *going to* option is more frequent in Jamaican English, and that this accords with a significantly greater use of progressives generally in that variety (1998: 209). She concludes that the difference between Jamaican English and international Standard English, in this respect, is "not manifested in the presence or absence of a feature, but in different usage preference patterns" (p. 212).

Similar findings have been reported by other researchers. Nelson (2003) studied the use of modal verbs expressing obligation and necessity (*must, should, ought to, need to, have to, have got to, gotta*) in six varieties of English (British, New Zealand, East Africa, India, Hong Kong, Jamaica). In the case of these modals, the usage preference patterns were distinctly different. While the most frequent modal in all varieties was found to be *have* (*got*) *to*, the other modals exhibit very different distribution patterns in the varieties under review. For instance, *need to* was found to be unusually frequent in Jamaican English (34 instances in 40,927 tokens), and unusually rare in Indian English (only one instance in 47,212 tokens) (Nelson, 2003: 28). Just as Sand found that the difference between varieties is not one of absence or presence, but of different distributions, the study of modals leads to a similar conclusion: in these terms at least, the difference is one of degree, not of kind.

Having said that, both Sand and Nelson discovered in their data examples of English usage which had not (as far as we know) been previously attested. In the Jamaican data, Sand found several instances of *have to be* VERB + *ing* with non-epistemic meaning:

. . . and we *have to be making* some new steel couplings to attach the new piece (Sand, 1998: 209–10)

Similarly, Nelson (2003: 31) found 57 instances of the following construction in the East African ICE corpus:

there *is need to* address cultural constraints

there *is need to* work on culture family and socialisation

This construction was found in both speech and writing, and in the data from both Tanzania and Kenya. It has not been found in any of the other currently available ICE corpora. It opens up many possible avenues of research. On the syntactic level, it is an interesting construction since the status of *need* is unclear. Is *need* verbal or nominal in this existential construction? Its nominal status is certainly unclear, since it never occurs (in this construction) with a determiner (*there is a need/there is the need*). On the pragmatic level, too, it opens up the whole question of when and why speakers would use an existential construction to express obligation.

More immediately, however, the discovery of these two constructions in ICE-Jamaica and in ICE-East Africa should lead us to revise our earlier observation. Perhaps some of the grammatical difference between varieties is, after all, a difference of kind, and not simply of degree. If this were proven to be true, it would inevitably force us to reconsider the notion of the "common core" which unites varieties of English. Findings such as these, together with those of Schneider (2000), cited above, offer a glimpse of the kinds of theoretical perspectives on world Englishes that become available for exploration using corpora. As Schneider puts it: "it is likely that at some point a larger set of such corpus-based results, drawn from further corpora and varieties will allow generalizations as to prototypical paths of linguistic evolution in New Varieties of English" (2000: 134).

5 The International Corpus of Learner English

Varieties of English from the Expanding Circle are catered for in the International Corpus of Learner English (ICLE), which is coordinated by Professor Sylviane Granger, University of Louvain-la-Neuve, Belgium. The ICLE project samples learner (EFL) English from a wide range of mother-tongue backgrounds, including French, German, Dutch, Spanish, Swedish, Finnish, Czech, Japanese, Chinese, Polish, and Russian (Granger, 1996). The ICLE corpus has been extensively used in studies of learner English, focusing, for example, on the forms of questions (Virtanen, 1998), the functions of participle clauses in academic writing (Granger, 1997), and the use of adverbial connectors (Granger and Tyson, 1996; Altenberg and Tapper, 1998).

6 Conclusion

As a discipline, corpus linguistics has come of age in recent years, but comparatively speaking, the corpus-based study of world Englishes is still in its

infancy. The collection and annotation of the ICE corpora has proved far more time-consuming than originally anticipated, despite the tireless efforts of researchers worldwide. A great deal of work remains to be done. On the practical side, the existing ICE corpora need to be completed. This will take place in at least two distinct phases: those corpora which have not been fully compiled in lexical form need to be completed, and then released for research. Following this, the corpora will have to be annotated to the level shown in ICE-GB, that is, with full part-of-speech tagging and syntactic analysis. There are also some major gaps in the geographical coverage that has been achieved so far. Most notably, very few African varieties are represented. Without proper

Table 41.3 An outline of the units and structures from morpheme to discourse which can be investigated using the completed ICE corpora (adapted from Kennedy, 1996: 223)

Word classes
  adjectives, adverbs, determiners, nouns, prepositions, pronouns, verbs, etc.
Word morphology and functions
  affixation, tense, number, etc.
Word types
Lemmas
Collocations
Phrases
  noun phrases, prepositional phrases, verb phrases, etc.
Clause elements
  subject, object, complement, adverbial, etc.
Clause patterns
  SV, SVO, SVOA, existential *there* constructions, etc.
Clause processes and information packaging
  extraposition, clefting, fronting, passivization, negation, etc.
Sentence types
  declarative, interrogative (*yes/no, wh-*), imperatives, etc.
Form and function
  interrogative versus question, etc.
Clause types
  subordinate clauses (nominal, relative, adverbial, comparative, etc.)
Clause relationships
  coordination, subordination, hypotaxis, parataxis
Discourse particles
Cohesion
Varieties and variation
  lexis, grammar, and discourse in different domains
    speech and writing
    sociolinguistic variation
    register variation
    regional variation

representation of important varieties from Nigeria, Cameroon, and Ghana – to mention just three – the ICE project will always offer only a partial picture of world Englishes.

On a more theoretical level, some consideration must be given to the methodology of comparing corpora. Previous work has been invaluable, though largely uncoordinated. Perhaps we now need a more coordinated approach, under the auspices of the ICE project. A useful starting point for such an approach is provided by Kennedy (1996), which offers an outline of the topics that can be investigated using the ICE corpora once they have been fully annotated to the same level as ICE-GB. Kennedy's outline is summarized in Table 41.3. As this table shows, the ICE corpora offer exciting possibilities for future research. This is especially true since most of the topics listed have never been systematically studied in most of the ICE varieties, and, for the most part, no comparative studies have ever been carried out on these topics. Though it is not exhaustive, Table 41.3 might be considered the starting point for a "prospectus" for future empirical research into world Englishes using the ICE corpora.

See also Chapters 20, WRITTEN LANGUAGE, STANDARD LANGUAGE, GLOBAL LANGUAGE; 35, A RECURRING DECIMAL: ENGLISH IN LANGUAGE POLICY AND PLANNING.

AVAILABILITY OF CORPORA

The corpora mentioned in this chapter are available as follows:

Brown, LOB, Kolhapur, ACE, Wellington:
  on a single CD-ROM from ICAME (International Computer Archive of Modern English), at the following address: The HIT Centre, Allegt. 27, N-5007, Bergen, Norway. Email: icame@hit.uib.no.
  Website: http://www.hd.uib.no.icame.html
  The Manuals for these corpora, cited in the references, are also available on the ICAME CD-ROM.

ICE-GB (The British component of the International Corpus of English):
  Survey of English Usage, University College London, Gower St, London WC1E 6BT, UK. Email: ucleseu@ucl.ac.uk.
  Website: http://www.ucl.ac.uk/english-usage/

ICE-New Zealand:
  Corpus Manager, Archive of New Zealand English, School of Linguistics and Applied Language Studies, Victoria University of Wellington, PO Box 600, Wellington, New Zealand.
  Website: www.vuw.ac.nz/lals/corpora/icenz.aspx

ICE-India, ICE-Singapore, ICE-East Africa, ICE-Philippines:
Dr Gerald Nelson, Department of English Language and Literature, University College London, Gower St, London WC1E 6BT, UK. Email: g.nelson@ucl.ac.uk.
Website: http://www.ucl.ac.uk/english-usage/ice/index.htm

ICLE corpora (components of the International Corpus of Learner English):
Professor Sylviane Granger, Centre for English Corpus Linguistics (CECL), Collège Erasme, Place Blaise Pascal 1, Université Catholique de Louvain, 1348 Louvain-la-Neuve, Belgium. Email: granger@lige.ucl.ac.be.
Website: http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/cecl.html

REFERENCES

Altenberg, Bengt (1998) *ICAME Bibliography 3 (1990–1998)*. http://www.hd.uib.no/icame.html.

Altenberg, Bengt and Tapper, Marie (1998) The use of adverbial connectors in advanced Swedish learners' written English. In *Learner English on Computer*. Edited by Sylviane Granger. London: Addison Wesley Longman, pp. 80–93.

Bansal, R. K. (1969) *The Intelligibility of Indian English*. Hyderabad: CIEFL.

Bauer, Laurie (1993) *Manual of Information to Accompany the Wellington Corpus of Written New Zealand English*. Wellington: Victoria University of Wellington.

Bauer, Laurie and Holmes, Janet (1996) Getting into a flap!: /t/ in New Zealand English. *World Englishes*, **15**(1), 115–24.

Bolt, Philip and Bolton, Kingsley (1996) The International Corpus of English in Hong Kong. In *Comparing English Worldwide: The International Corpus of English*. Edited by Sidney Greenbaum. Oxford: Clarendon Press, pp. 197–214.

Collins, Peter (1988) The semantics of some modals in contemporary Australian English. *Australian Journal of Linguistics*, **8**, 233–58.

Collins, Peter (1991) The modals of obligation and necessity in Australian English. In *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. Edited by Karin Aijmer and Bengt Altenberg. London: Longman, pp. 145–65.

Fallon, Helen (2004) Comparing world Englishes: A research guide. *World Englishes*, **23**(2), 309–16.

Francis, W. Nelson and Kučera, Henry (1971) *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*. Revised edition. Providence: Department of Linguistics, Brown University.

Glauser, Beat, Schneider, Edgar W., and Görlach, Manfred (1993) *A New Bibliography of Writings on Varieties of English 1984–1992/3*. Amsterdam: Benjamins.

Granger, Sylviane (1996) Learner English around the world. In *Comparing English Worldwide: The International Corpus of English*. Edited by Sidney Greenbaum. Oxford: Clarendon Press, pp. 13–24.

Granger, Sylviane (1997) On identifying the syntactic and discourse features of participle clauses in academic English: Native and non-native writers compared. In *Studies in English Language and Teaching in Honour of Flor Aarts*. Edited by Igne de Mönnink Aarts and Herman Wekker. Amsterdam: Rodopi, pp. 185–98.

Granger, Sylviane and Tyson, Stephanie (1996) Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes*, **15**(1), 17–27.

Greenbaum, Sidney (1988) A proposal for an International Corpus of English. *World Englishes*, **7**, 315.

Greenbaum, Sidney (1993) The tagset for the International Corpus of English. In *Corpus-Based Computational Linguistic*. Edited by Clive Souter and Eric Atwell. Amsterdam: Rodopi, pp. 11–24.

Greenbaum, Sidney (1996) *Oxford English Grammar*. Oxford: Oxford University Press.

Greenbaum, Sidney and Nelson, Gerald (1995a) Clause relationships in spoken and written English. *Functions of Language*, **2**, 1–21.

Greenbaum, Sidney and Nelson, Gerald (1995b) Nuclear and peripheral clauses in speech and writing. In *Studies in Anglistics*. Edited by Gunnel Melchers and Beatrice Warren. Stockholm: Almqvist and Wiksell, pp. 181–90.

Greenbaum, Sidney and Nelson, Gerald (1996) Positions of adverbial clauses in British English. *World Englishes*, **15**(1), 69–81.

Greenbaum, Sidney, Nelson, Gerald, and Weitzman, Michael (1995) Complement clauses in English. In *Using Corpora for Language Research: Studies in Honour of Geoffrey Leech*. Edited by Jenny Thomas and Mike Short. London: Addison Wesley Longman, pp. 76–91.

Holmes, Janet (1993) Sex-marking suffixes in written New Zealand English. *American Speech*, **68**, 357–70.

Holmes, Janet (1996) The New Zealand spoken component of ICE: Some methodological challenges. In *Comparing English Worldwide: The International Corpus of English*. Edited by Sidney Greenbaum. Oxford: Clarendon Press, pp. 163–81.

Holmes, Janet, Vine, Bernadette, and Johnson, Gary (1998) *Guide to the Wellington Corpus of Spoken New Zealand English*. Victoria University of Wellington: School of Linguistics and Applied Language Studies.

Johansson, Stig and Hofland, Knut (1989) *Frequency Analysis of English Vocabulary and Grammar: Based on the LOB Corpus*. 2 vols. Oxford: Oxford University Press.

Johansson, Stig, Leech, Geoffrey, and Goodluck, Helen (1978) *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*. Oslo: University of Oslo.

Kachru, Braj B. (1965) The Indianness in Indian English. *Word*, **2**, 391–410.

Kachru, Braj B. (1985) Standards, codification and sociolinguistic realism: The English language in the outer circle. In *English in the World: Teaching and Learning the Language and Literatures*. Edited by Randolph Quirk and Henry G. Widdowson. Cambridge: Cambridge University Press, pp. 11–30.

Kennedy, Graeme (1996) The corpus as a research domain. In *Comparing English Worldwide: The International Corpus of English*. Edited by Sidney Greenbaum. Oxford: Clarendon Press, pp. 217–26.

Kennedy, Graeme (1998) *An Introduction to Corpus Linguistics*. London: Longman.

Kučera, Henry and Francis, W. Nelson (1967) *Computational Analysis of Present-Day American English*. Providence: Brown University Press.

Leitner, Gerhard (1992) *Begin* and *start* in British, American and Indian English. *Hermes*, **13**, 99–122.

Leitner, Gerhard and Hesselmann, M. (1996) "What do you do with a ball in soccer?" Medium, mode, and pluricentricity in soccer reporting. *World Englishes*, **15**(1), 83–102.

McEnery, Tony and Wilson, Andrew (2001) *Corpus Linguistics*. 2nd edition. Edinburgh: Edinburgh University Press.

Meyer, Charles F. (1996) Coordinate structures in English. *World Englishes*, **15**(1), 29–41.

Nelson, Gerald (1996a) The design of the corpus. In *Comparing English Worldwide: The International Corpus of English*. Edited by Sidney Greenbaum. Oxford: Clarendon Press, pp. 27–35.

Nelson, Gerald (1996b) Markup systems. In *Comparing English Worldwide: The International Corpus of English*. Edited by Sidney Greenbaum. Oxford: Clarendon Press, pp. 36–53.

Nelson, Gerald (2003) Modals of obligation and necessity in varieties of English. In *From Local to Global English: Proceedings of the Style Council 2001/2*. Edited by Pam H. Peters. Sydney: Dictionary Research Centre, Macquarie University, pp. 25–32.

Nelson, Gerald, Wallis, Sean, and Aarts, Bas (2002) *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: Rodopi.

Peppé, Sue (1995) The Survey of English Usage and the London-Lund Corpus: Computerizing manual prosodic transcription. In *Spoken English on Computer: Transcription, Mark-up and Application*. Edited by Geoffrey Leech, Greeg Myers, and Jenny Thomas. London: Longman, pp. 187–202.

Peters, Pam (1993a) American and British English in Australian usage. In *Style on the Move: Proceedings of the Style Council 92*. Edited by Pam Peters. Sydney: Dictionary Research Centre, Macquarie University, pp. 20–7.

Peters, Pam (1993b) Corpus evidence on some points of usage. In *English Language Corpora: Design, Analysis and Exploitation*. Edited by Jan Aarts, Pieter de Haan, and Nelleke Oostdijk. Amsterdam: Rodopi, pp. 247–55.

Peters, Pam (1994) American and British influence in Australian verb morphology. In *Creating and Using English Language Corpora*. Edited by Udo Fries, Gunnel Tottie, and Peter Schneider. Amsterdam: Rodopi, pp. 149–58.

Peters, Pam, Collins, Peter, Blair, David, and Brierley, Alison (1988) The Australian corpus project: Findings on some functional variants in the Australian press. *Australian Review of Applied Linguistics*, **11**(1), 22–33.

Quirk, Randolph, Leech, Geoffrey, Svartvik, Jan, and Greenbaum, Sidney (1985) *A Comprehensive Grammar of the English Language*. London: Longman.

Sand, Andrea (1998) First findings from ICE-Jamaica: The verb phrase. In *Explorations in Corpus Linguistics*. Edited by Antoinette Renouf. Amsterdam: Rodopi, pp. 201–16.

Sayder, S. (1989) The subjunctive in Indian, British, and American English: A corpus-based study. In *Englische Textlinguistik und Varietätenforschung*. Edited by Gottfried Graustein and Wolfgang Thiele. Leipzig: Karl-Marx-Universität, pp. 58–66.

Schmied, Josef (1990) Corpus linguistics and non-native varieties of English. *World Englishes*, **9**(3), 255–68.

Schmied, Josef (1995) National standards and the International Corpus of English. In *New Englishes: A West African Perspective*. Edited by Ayọ Bamgbọse, Ayọ Banjo, and Andrew Thomas. Ibadan, Nigeria: Mosuro, pp. 337–48.

Schmied, Josef and Hudson-Ettle, Diana (1996) Analyzing the style of East African newspapers in English. *World Englishes*, **15**(1), 103–13.

Schneider, Edgar W. (2000) Corpus linguistics in the Asian context: Exemplary analyses of the Kolhapur corpus of Indian English. In *Parangalcang Brother Andrew: Festschrift for Andrew Gonzalez on His Sixtieth Birthday*. Edited by Ma. Lourdes S. Bautista, Teodora A. Llamzon, and Bonifacio P. Sibayan. Manila: De La Salle University Press, pp. 115–37.

Shastri, S. V. (1986) *Manual of Information to Accompany the Kolhapur Corpus of Indian English, for Use with Digital Computers*. Kolhapur: Shivaji University.

Shastri, S. V. (1988) The Kolhapur corpus of Indian English and work done on its basis so far. *ICAME Journal*, **12**, 15–26.

Shastri, S. V. (1992) Opaque and transparent features of Indian English. In *New Directions in English Language Corpora: Methodology, Results, Software Developments*. Edited by Gerhard Leitner. Berlin/New York: Mouton de Gruyter, pp. 263–75.

Sigley, Robert (1997) The influence of formality and channel on relative pronoun choice in New Zealand English. *English Language and Linguistics*, **1**(2), 207–32.

Van Halteren, Hans and Oostdijk, Nelleke (1993) Towards a syntactic database: The TOSCA analysis system. In *English Language Corpora: Design, Analysis and Exploitation*. Edited by Jan Aarts, Pieter de Haan, and Nelleke Oostdijk. Amsterdam: Rodopi, pp. 145–61.

Virtanen, T. (1998) Direct questions in argumentative student writing. In *Learner English on Computer*. Edited by Sylviane Granger. London: Addison Wesley Longman, pp. 94–110.

Yang, Wen (1997) Discourse analysis of direct and indirect speech in spoken New Zealand English. *New Zealand Studies in Applied Linguistics*, **3**, 62–78.

FURTHER READING

Biber, Douglas (1993) Representativeness in corpus design. *Literary and Linguistic Computing*, **8**, 1–15.

Bolton, Kingsley (ed.) (2002) *Hong Kong English: Autonomy and Creativity*. Hong Kong: Hong Kong University Press.

Chafe, Wallace (1992) The importance of corpus linguistics to understanding the nature of language. In *Directions in Corpus Linguistics*. Edited by Jan Svartvik. Berlin: Mouton de Gruyter, pp. 79–97.

Granger, Sylviane (ed.) (1998) *Learner English on Computer*. London: Addison Wesley Longman.

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

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

Hundt, Marianne (1998) *New Zealand English Grammar: Fact or Fiction?* Amsterdam: Benjamins.

Mair, Christian (1992) Problems in the compilation of a corpus of standard Caribbean English: A pilot study. In *New Directions in English Language Corpora: Methodology, Results, Software Developments.* Edited by Gerhard Leitner. Berlin/New York: Mouton de Gruyter, pp. 75–96.

Meyer, Charles F. (2002) *English Corpus Linguistics: An Introduction.* Cambridge: Cambridge University Press.

Oostdijk, Nelleke (1991) *Corpus Linguistics and the Automatic Analysis of English.* Amsterdam: Rodopi.

Peters, Pam (1998) In quest of international English: Mapping the levels of regional divergence. In *Explorations in Corpus Linguistics.* Edited by Antoinette Renouf. Amsterdam: Rodopi, pp. 281–92.

Quirk, Randolph (1992) On corpus principles and design. In *Explorations in Corpus Linguistics: Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991.* Edited by Jan Svartvik. Berlin: Mouton de Gruyter, pp. 457–69.

Schmied, Josef (1991) *English in Africa: An Introduction.* London/New York: Longman.

Schneider, Edgar W. (2003) Evolution(s) in global English(es). In *From Local to Global English: Proceedings of the Style Council 2001/2.* Edited by Pam H. Peters. Sydney: Dictionary Research Centre, Macquarie University, pp. 3–24.