Rassegna Italiana di Linguistica Applicata

ESTRATTO

Anno XLIII

Gennaio-Agosto 2011 / 1-2
ISSN 0033-9725

USING CORPORA AS AN INNOVATIVE TOOL TO COMPARE VARIETIES OF ENGLISH AROUND THE WORLD: THE *INTERNATIONAL CORPUS OF ENGLISH*[*]

JOSEF SCHMIED
Chemnitz University of Technology

Abstract

*L'articolo ripercorre, in chiave personale e programmatica, i primi venti anni dell'*International Corpus of English. *Il resoconto volge, da un lato, lo sguardo all'indietro, ricordando le fasi iniziali del progetto e soffermandosi sul lavoro svolto dai numerosi gruppi di ricerca che vi hanno contribuito, e dall'altro tenta, programmaticamente, di disegnare i percorsi di sviluppo futuri e di individuare nuove strade di ricerca su base comparativa. L'articolo propone inoltre un piccolo studio sulla modalità, al fine di mostrare come l'utilizzo dell'*International Corpus of English *permetta di gettare luce nuova su temi che da sempre interessano la ricerca linguistica.*

* A related research proposal was later submitted as part of the Special Research Programme (SFB) "Identity in Africa" to the German Research Foundation (DFG) and supported between 1989 and 1996 so that the compilation and publication of the corpus with accompanying manual could basically be finished. I wish to thank my long-term ICE partners, esp. Gerry Nelson, my project assistants Diana Hudson Ettle, Eva Hertel and Barbara Kauper, and my later colleagues Christoph Haase and Susanne Wagner for many interesting discussions.

1. *The (hi)story of ICE*

1.1 *Beginnings and concepts*

The beginnings of the project *International Corpus of English* (henceforth *ICE*) can be traced back to the ICAME (International Computer Archive of Modern English) Conference in Birmingham in May 1988. I gave a paper on "Compiling a corpus of East African English" (Schmied 1989), which discussed basically in what way the categories of the Brown, LOB and Kolhapur models can be applied to second language varieties in Africa. The essence of the presentation was that in second language communities, i.e. most varieties of English in Africa, sociolinguistic variables like gender, status, age and particularly first language and education played a dominant role in stratifying a "national corpus" which was to serve as a basis for intra- and international comparisons. Sidney Greenbaum had just written "A proposal for an international computerised corpus of English", in which he suggested (Greenbaum 1988:31):

1. to sample standard varieties from other countries where English is the first language, for example Canada and Australia;
2. to sample national varieties from countries where English is an official additional language, for instance India and Nigeria; and
3. to include spoken and manuscript English as well as printed English.

Thus the ICE project expanded corpus linguistics from the mother-tongue varieties to all "Anglophone countries" world-wide. From my perspective, this was a major achievement since it meant that for the first time the sociolinguistic stratification of English language variation was seriously discussed in corpus compilation. This included the "New Englishes" discourse, which had been started by Braj Kachru only a few years before (as summarized in Kachru 1990). These discussions also determined the borderline of the ICE project, which was to include second-language varieties but not foreign-language varieties. This was to be left to a parallel project, the *International Corpus of Learner English* (*ICLE*[1] cf. Granger 1998), which emphasized mother-tongues and neglected text-type stratification. If we look at conferences on "World Englishes" and

[1] The development of ICLE can be followed on its website (http://www.uclouvain.be/en-cecl-icle.html). It will not be discussed here since it is well documented there and separated from ICE over a decade ago.

"Learner English" today, the impact of corpus-linguistic methodology on these fields is impressive.

1.2 *ICE problems*

Besides funding, which was never centrally applied for and discussed in ICE, numerous problems have prevented the individual ICE projects from completing their work in the 1990s, as originally envisaged. Although initially Sidney Greenbaum coordinated the project especially through regular ICE meetings during the ICAME conferences that most ICE teams attended and the ICE circulars afterwards, many problems at national level could not be solved even with his support through letters and discussions. From the very start, ICE-GB was seen by Greenbaum as a model corpus (also for grammatical description; cf. Greenbaum 1996), but it also has its limitations. The major problem may be availability, because it is not freely available to *bona-fide* researchers like the other ICE corpora, partly because it is distributed with ICE-CUP, the "Computer Utility Program" that was specially developed for grammatical analyses to allow POS and even tree searches. The sociolinguistic problem with ICE-GB is that it is clearly centered on the Survey of English Usage at University College, London; it does not adequately cover other regions of England, to say nothing of Scotland. Of course, Northern Ireland is covered in ICE-Ireland. Since the British National Corpus (BNC) has become available, it has clearly been used as *the* reference corpus for Great Britain, since it covers similar text-types and a similar period (the early 1990s) and is nowadays easily accessible through various web interfaces. This means in practice, since ICE-New Zealand and ICE-Australia are not freely available as a simple down-load from the WWW either, that the only native-speaker corpora available at the moment are ICE-Ireland and ICE-Canada.

Unfortunately, major 'Anglophone countries' like the US and Nigeria have not been able to compile a 'national corpus' despite the attempts of many linguists[2]. Some projects had to wait until the political (and funding) climate became more positive (like ICE Ireland, which was published together with an exemplary User Guide in 2008). Many 'third-world' projects had to rely on the collaboration with European, especially German linguists (e.g. ICE Ghana or ICE Nigeria). In this context, only two general issues can be discussed in an exemplary form: the tension

[2] A simple alternative nowadays are the BYU-Corpora compiled by Mark Davies, which even surpass the BNC in size and historical depth (http://corpus.byu.edu/).

between culture-specific and standardized corpus compilation and the issue of distributing raw vs. processed corpora.

1.2.1 *Adaptation to sociolinguistic and cultural contexts*

From the beginning of ICE, a major discussion has always been the balance between representativeness vs comparability (Schmied 1990). To what extent should the sampling procedures be adapted to the sociolinguistic and cultural background when, for instance, certain text-types played a more or less important role in a specific national context? Usage patterns depend very much on the cultural context, so that social letters were much more frequent in the first world than in the third world in the 1990s and have largely been replaced with email where the internet is accessible easily or by mobile phones and texting in other parts of the world. The idea of taping private conversations is as alien to some cultures as that of recording legal proceedings to others. The copyright restrictions may not always be clear enough, but they have always been a major problem for the distribution of ICE corpora even to *bona-fide* researchers, as witnessed by e.g. special copyright agreement forms to be signed when obtaining ICE-Ireland or ICE-NZ.

Thus the Tanzanian component of ICE-East Africa is smaller than the Kenyan component, since the functions of English in some spoken registers are limited (with almost 290,000 spoken words in ICE-Kenya vs only 214,000 spoken words in ICE-Tanzania; cf. Hudson Ettle, Schmied 1999). This, however, makes ICE-Tanzania and ICE-Kenya not directly compatible with each other or the other ICE corpora. The simple fact that all frequency counts have to be normalised to 1 million words or occurrence per 100,000 words has considerably restricted its use for quick corpus comparisons.

1.2.2 *Corpus development: tagging and parsing*

Corpus annotation and processing has been a major issue in early ICE discussions. ICE-GB received several national grants to develop the software for ICE corpora, especially ICE-GB, but the other ICE corpora have profited relatively little from these efforts, despite a lot of personal commitment by ICE-GB members or the current ICE coordinator Gerry Nelson over the last few years (with ICE-Philippines, for instance). Since all ICE teams are responsible for their own annotation, the following examples are taken from ICE-GB.

ICE-GB texts were automatically tagged for Part-of-Speech (POS) or word-class by the ICE Tagger developed by Sean Wallis at the Survey of English Usage, University College London. The tagger assigns word-class tags to each lexical item in the corpus. The tag-set had been developed especially for ICE and is largely based on Quirk et al.'s (1985) *A Comprehensive Grammar of the English Language* and exemplified in Nelson, Wallis, Aarts (2002). The following examples are also available at the ICE website (http://ice-corpora.net/ice/annotate.htm):

Each	PRON(univ,sing)
of	PREP(ge)
these	PRON(dem, plu)
is	V(cop, pres)
the	ART(def)
responsibility	N (com,sing)
of	PREP(ge)
one	NUM(card, sing)
person	N(com, sing)

On this basis, every sentence in the corpus is analysed at phrase, clause, and sentence level, and the analysis can be shown in the form of a parsed tree (Fig. 1).

The unique advantage of ICE-GB is that ICE-CUP can be used for syntactic searches and comparisons. The disadvantage of the system is that checking the POS tagging (despite an accuracy of over 90 percent) is simply too labour-intensive for most ICE projects since it requires the expertise of postgraduate students. This means that if syntactic questions are not central to the national ICE projects, they are often postponed; and if they are too (lexico-)specific, the relatively small size of ICE corpora makes comparisons relatively difficult (e.g. Schneider 2004 for particle verbs).

2. *The present status of ICE*

The current status of the International Corpus of English can be followed on its website (http://ice-corpora.net/ice/). This also gives the latest news (in mid-2010), e.g. that ICE-Bahamas has joint the ICE family

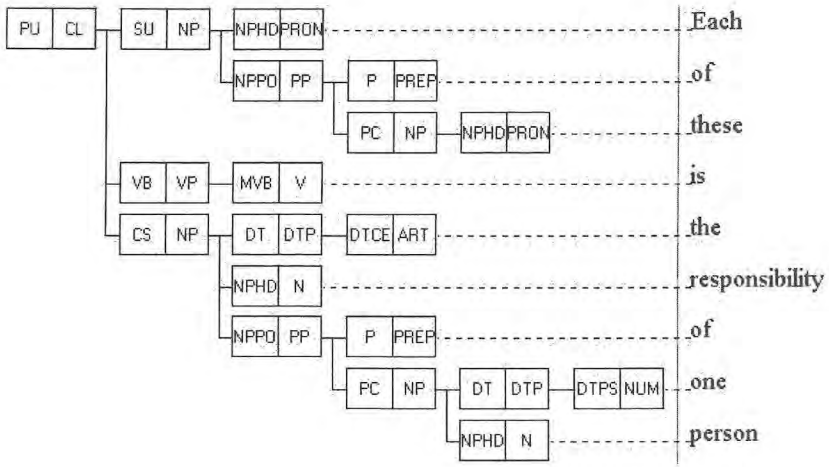

Figure 1. *Parsed tree for* Each of these is the responsibility of one person

recently and that ICE-Canada has been released recently. It lists the corpora currently available freely to *bona-fide* researchers who sign a licence agreement which grants that the texts are only used for scholarly purposes.

This page shows that ten ICE corpora are available, the first language corpora from Canada, Great Britain, Ireland, and New Zealand, and the second-language corpora from East Africa (i.e. Kenya and Tanzania), India, Jamaica, Hong Kong, the Philippines and Singapore. Thus, interestingly enough the real value for the ICE project today is that it makes available lesser-known varieties from Africa and Asia. But the website also shows that many more corpora are still being compiled from Australia, the Bahamas, Fiji, Ghana, Malaysia, Malta, Namibia, Nigeria, Pakistan, South Africa, Sri Lanka, Trinidad & Tobago and the USA.

The current ICE association is very loose. Its members meet at regular international conferences like ICAME or IAWE for workshops with progress reports and presentations. The technical support that Sidney Greenbaum used to provide is difficult to maintain for the Survey of English Usage and the ICE coordinator can usually only provide limited support for funding applications and small workshops.

As always, today new teams can join easily and are encouraged to use the standard ICE principles as they are presented on the website. Culture- or media-specific adaptations, as discussed above, are usually

supported by the other ICE teams, although they are aware of the problem of comparability.

There are few ICE publications where many ICE partners publish together. Apart from the first ICE book edited by the founding coordinator Sidney Greenbaum in 1996, which covers a wide spectrum of problems of compilation and annotation, implementation and applications, only two special issues of *World Englishes*, one edited by Greenbaum and Nelson in 1996 and one edited by the current coordinator Gerald Nelson in 2004 document the development of ICE mainly through case studies.

Recently a new discussion opportunity has been established with the e-journal *ICES – International Corpus of English Studies*, whose major aim is to publish the ICE discussions and workshop presentations together with other comparative studies that may serve as inspiration for the further discussion of English world-wide.

This does not mean that individual ICE researchers and ICE teams have not made further contributions to methodological or theoretical, descriptive or applied issues, but these were usually not made on an explicit comparative ICE basis.

3. *ICE case studies: modalities*

The following case study exemplifies how our comparative study of ICE corpora can shed the light on old practical and theoretical questions. It uses modal auxiliaries, because modality is a decisive feature of English that displays interesting differences between first and second language varieties. The following analysis is undertaken to investigate four different hypotheses:

1. English auxiliaries are very unevenly distributed in actual language usage.
2. The meanings of modal auxiliaries are so different that epistemic and deontic use have to be distinguished; even though epistemic use is often less prominent in second language varieties, it is usually more frequent than the original deontic use.
3. The use of modal auxiliaries and the proportion of epistemic and deontic use in particular are very culture-specific and thus vary greatly between ICE varieties.
4. This variation can be explained on the basis of the sociolinguistic data of English in the respective ICE community.

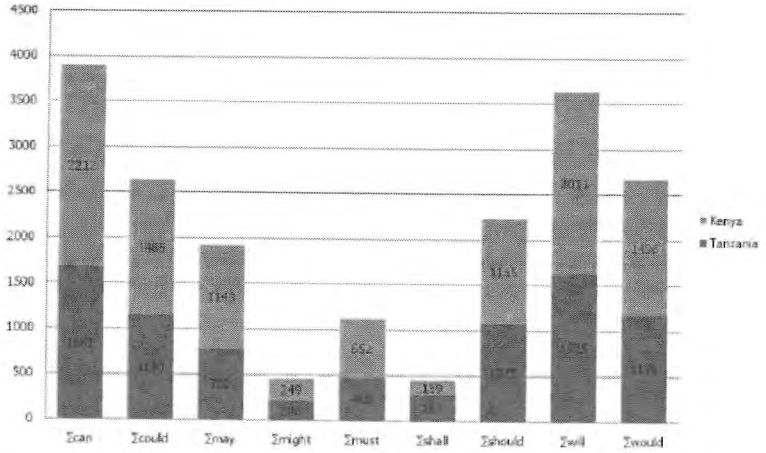

Figure 2. *The distribution of central modal auxiliaries in Kenya and Tanzania (ICE-EA)*

A detailed analysis of the actual usage of modal auxiliaries in the two national parts of ICE-East Africa, Kenya and Tanzania is particularly useful since the collection principles and corpus processing were obviously very similar, as these texts were compiled by the same ICE team and the underlying sociolinguistic differences are well documented, for instance in the accompanying Manual (Hudson Ettle, Schmied 1999).

Figure 2 shows the distribution of central modal auxiliaries in Kenya and Tanzania from ICE-East Africa. As the columns clearly prove, the use of *can* is more than ten times as frequent as the use of *shall* or *might*. With the exception of *shall/should*, the use of the past tense form is always less frequent than the base form (i.e. *can* > *could, may* > *might, will* > *would*). This does not mean, however, that past tense meanings are more frequent, as the columns of Figure 3 clearly show that the epistemic usage is more prominent than the deontic – with the exceptions of *must*, *shall* and *would*.

Of course, isolated comparisons of modal auxiliaries do not mean very much – they have to be compared to other ICE corpora. Comparative data made available from other ICE teams (Table 1) show that all modal auxiliaries are used less often in Kenya and Tanzania than in other 'Anglophone nations' (Tab. 1).

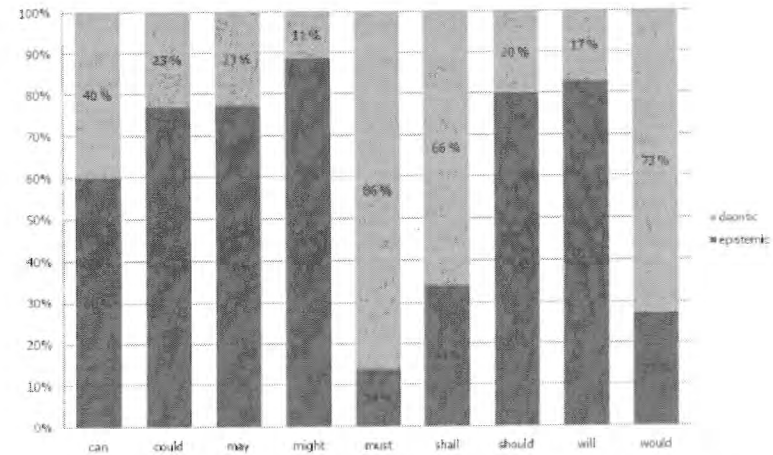

Figure 3. *Deontic and epistemic usage of central modals in ICE-EA*

Modal auxiliary	*Total per corpus*				*Per million words*			
	ICE-GB	ICE-Phil/10P	ICE-K	ICE-T	ICE-GB	ICE-Phil*	ICE-K	ICE-T
can	3574	425	2212	1681	3574	3601	1580	1201
could	1635	130	1485	1150	1635	1101	1061	821
may	1219	120	1143	782	1219	1016	816	559
might	693	45	249	208	693	381	178	149
must	687	55	652	468	687	466	466	334
shall	222	30	159	287	222	254	114	205
should	1117	100	1155	1075	1117	847	825	768
will	2841	505	2011	1628	2841	4279	1436	1163
would	3037	270	1496	1176	3037	2288	1069	840
Total	15025	1680	10562	8455	15025	14237	7545	6040
* normalised from 10% sample								

Table 1. *Modal auxiliaries in ICE-GB, ICE-Phil and ICE-EA (-K/-T)*

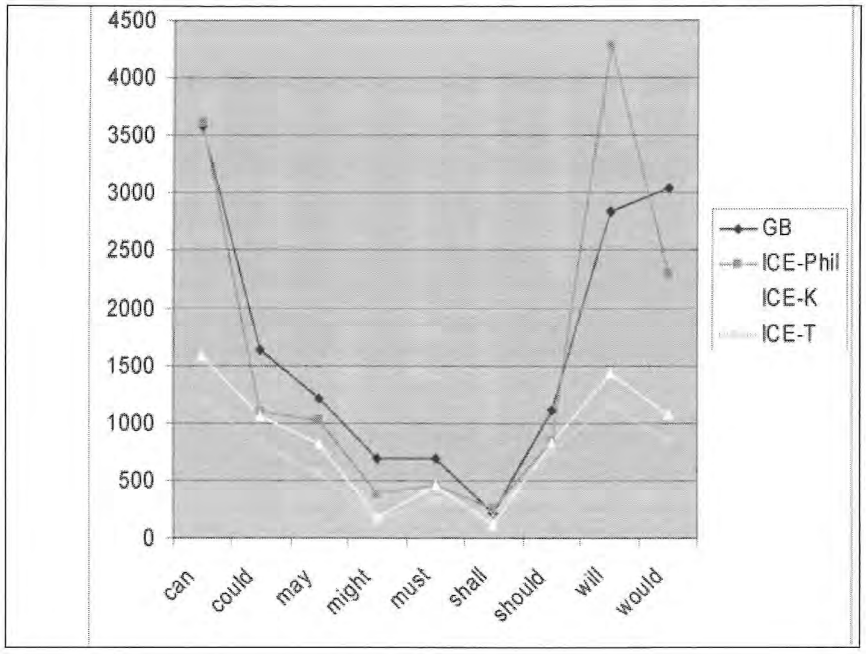

Figure 4. *Central modal auxiliaries in ICE-GB, -Philippines, -Kenya and -Tanzania*

The comparison here requires some adaptation, since the basis of analysis is not the same 1 million word corpus. The ICE-Philippines' figure is based on a sample and has to be multiplied accordingly. The figures from Kenya and Tanzania have to be 'normalised' to their occurrence per 1 million or 1000,000 words.

In a different presentation (Figure 4), the parallel lines from Kenya and Tanzania are even more obvious. Central modals are obviously used more frequently in Great Britain and the Philippines; especially *can* and *will/would* deserve a more detailed comparative analysis.

This small case study shows that a bottom-up approach that starts with examples from a corpus shows interesting differences between speech communities. From a more theoretical perspective, the main hypothesis has been confirmed that through a higher frequency of modal auxiliaries in general and of epistemic use in particular, English in Kenya shows that it has developed further towards a New English variety than English in Tanzania. The few exceptions in the standard pattern are interesting and can be explained again on the basis of the sociolinguistic status of English in the two East African countries. Since English in Tanzania is the more

formal variety, as informal texts are expressed more often in Kiswahili, the most formal modal auxiliary *shall* occurs more frequently than expected.

Further analyses are, however, necessary since the simple distinction between epistemic and deontic use is not always enough, e.g. in lexeme-specific cases like the habitual and historical uses of *would*.

4. *The future of ICE (ICE2)*

4.1 *Changes in international communication since 1989*

Since the first ICE discussions in 1989, we have had a revolution in international communication which is part of what we call globalization today. On the one hand, traditional forms of communication have been replaced (e.g. 'mail' has been replaced with 'email'). On the other hand, the world-wide web provides us with new genres like personal webpages, blogs and chats, which expand the traditional text types that were discussed at the beginning of ICE. The question today is in what way it is enough to replace old text types by new ones and in what way the whole spectrum has to be reconsidered.

At the same time, the discussion of English as a Global Language has given a new boost to the concept of English as an International Language (EIL) or English as a Lingua Franca (ELF), within the European Union in particular since the expansion to Central and Eastern Europe (Schmied 2010). Another major example is China, although relatively few people regularly speak and use English in China today. The fact that almost everyone aims at learning English will make Chinese English one of the major varieties of the future.

Within the ICE discussions, this sheds new light on the concept of English-medium education, because many more countries consider secondary and tertiary education through the medium of English as essential. A major restriction of ICE during the initial discussions was that an "Anglophone country" was defined as one having at least four years of English-medium education up to the age of eighteen. If this requirement is dropped, this expands the possible members of the ICE community considerably.

4.2 *New methods of corpus compilation*

In corpus compilation, the one-million word corpus, which had been a standard over 20 years ago, has been replaced with attempts to create

new national corpora, which follow basically the model of the British National Corpus (BNC, 1991-93). Although the corpus of American English has made very little progress over the last decade, new national attempts such as in Australia have made larger comparative corpora of about one hundred million words a new norm.

Thus, the comparative expansion of ICE could go in two directions. The diachronic changes could be followed up by new ICE corpora that record compatible text types 20 to 25 years later. If we consider the early 1990s as the starting point, the first language corpora (Great Britain and New Zealand) and the second language corpora (East Africa – Kenya/Tanzania) have been compiled relatively early. A new conservative re-launch of ICE would be relatively simple if only the old-fashioned text types like social letters are adapted. The second direction is the compilation of a web-based monitor corpus that opens new possibilities in particular for the less frequent and more collocational features of the English language that a one-million-word corpus is hardly sufficient for.

Such a web-monitor corpus could be extracted from the internet through a keyword approach, in which business letters, for instance, could be identified through the appropriate discourse-initial or -final keywords (e.g. *Yours sincerely*) or administrative writing could be retrieved by searching for *regulations* or even *FAQ*. On the basis of the standard web search engines, special English language websites could be identified in each domain. To these website copiers like HTtrack could be applied and large raw data files could be copied literally overnight. The cleaning of these files and the text-specific tagging would still be a major undertaking if we insisted on a relatively 'clean' copy, but this procedure would give us a workable monitor corpus relatively quickly (Nelson 2010).

In this way, a quick and dirty 'national corpus' could be compiled almost overnight. The Chinese varieties of English are an interesting example for the current debate. Whereas the Hong Kong and Singapore varieties have been analysed, the mainland China variety that is based on the standard Chinese Putonghua has attracted much less attention, although there is a widely accepted model at least for the written standard in the form of the well-known newspaper *China Daily*. This is one reason why we have, for instance, established a big and dirty corpus of 'China Daily English', which comprises some 800,000 texts with some 40 million words (which shows that the average of China Daily texts is surprisingly small). In China, this newspaper is unrivalled and thus national decisions of usage norms can be easily and acceptably based on such a corpus.

4.3 *Applied issues: towards new national norms?*

Comparative analyses will also make it easier for the responsible political and academic agencies or institutions to acknowledge new norms. According to Kachru's (1990) model of concentric circles, the outer circle is characterized as "norm-developing", but this is a major task for New Englishes. In countries like Kenya or Tanzania, a new national examination board has been set up to replace the colonial Cambridge-based correction of nation-wide school-leaving or university entrance exams. These bodies need guidelines for the acknowledgement of new usage norms in their country and beyond. If they see that new patterns emerge not only in their national variety of English but also in neighbouring countries or countries where English plays a similar role as a second language, they may be more inclined to accept them than when they seem idiosyncratic to certain speakers or speaker groups (usually from certain educational levels).

The usage of certain idiomatic expressions may be less fixed in New Englishes than in native English varieties with a long normative tradition. The *grassroots* examples (Table 2) show despite relatively small figures that *grassroots* are discussed more often in Tanzania than in Kenya, which seems plausible since Tanzania has a much more open development debate than Kenya. However, the orthographic and morphological conventions are not always easy to analyse. The spoken component of any corpus depends very much on the transcription, which depends on individual transcribers (despite the general guidelines about pauses). In addition to these transcription and pronunciation conventions, we also have to consider the wider context since the premodifying form (as in *grassroot levels*) seems to be used less often than expected. This, however, depends very much on the syntagmatic context. In the case of ICE East Africa out of the 41 *grassroots* occurrences only five occur in a premodifying position.

	Kenya			Tanzania			Σ
	Σ	*written*	*spoken*	Σ	*written*	*spoken*	
grassroots	4	3	1	16	10	6	20
grass roots	12	11	1	9	3	6	21
grass root		1				1	2
Σ	16			25			41

Table 2. *Writing conventions of* grassroot(s) *in ICE-East Africa*

The empirical analysis of ICE corpora is a first step towards the establishment of a grammatical and a lexical database which may in the end lead to national grammars and dictionaries that could be used for in-group and out-group reference.

A relatively clear case of national usage norms are cultural lexemes, like East Africanisms. A standard example that is not as widely known outside of Kenya and Tanzania as *askari* or *safari* (two lexemes that have made it relatively early into the *Oxford English Dictionary*) is the term *matatu*, which could have a dictionary entry based on ICE-EA sentences like:

> ***matatu*** pl ˜*s N* 'collective taxi' in EAfr., esp. Kenya
> usu. Licensed for fixed routes of public transport, but flexible, they leave when 'full'; infamous for reckless driving and overcrowding;
> *etym.* <Sw, three, *orig.* 3 Shs fare;
> *coll.*: *agent* driver, tout, operator, passenger; *loc.* park, stand, stage, stop
> *P* in, on board a ˜; *V* enter, board

In the end, this database norm approach may lead to new norms that replace the traditional native speaker from Britain, which is still seen as the norm by most African countries. In South East and East Asia, the norm tends much more towards the American model. The comparison of ICE corpora helps to pursue these trends and add a descriptive basis to the often impressionistic evidence.

4.4 *Theoretical issues: dynamic model – new categorisation?*

A comparative approach may also be useful for developing further the theoretical modelling of new English developments. In this debate, Schneider (2006) has proposed the following evolutionary stages: from the foundation of a new variety the exonormative stabilisation and the sociolinguistic nativisation may lead to endonormative stabilisation and finally differentiation. Since most of the stages are gradual, a corpus-linguistic approach is, of course, appropriate since it allows a probabilistic component in a usage analysis. It is also questionable whether an

endonormative specialisation occurs before stylistic differentiation; both are based on a new cognitive perception that several structures are possible and plausible, and less normative attitudes towards certain forms and patterns thus emerge.

Similar 'developmental' issues have been brought up recently by Mesthrie and Bhatt (2008:90):

> One such broad dichotomy involves varieties that favour deletion of elements and those that disfavour it. In this regard the differences between Sgp Eng (especially amongst those with Chinese substrates) and African varieties are striking.

They (ibid:91) propose three examples as distinctive:

Come what may (come).
He made me (to) do it.
As you know (that) I am from the Ciskei.

and they call these varieties "deleters vs preservers" and suggest that East Asian New Englishes may be the former and the (East) African varieties the latter. A different hypothesis could be that the more integrated a variety is within a speech community the more deletion is tolerated, and the less integrated it is the more structures have to be preserved to make meanings more transparent to users at lower stages of language learning. In this argumentation, there would be more deletions in ICE Singapore than in ICE Malaysia or in ICE Kenya than in ICE Tanzania. Unfortunately, other ICE varieties are not yet completely available to set up theoretical gradients in West and East Africa as well as, for instance, in South East Asia.

5. *Conclusion*

So far in Corpus Studies we have two major types of corpora: we have the diachronic, small but tidy family of the BROWN corpora from the major referent varieties, American and British English, that allow us to follow the development of Present Day English over almost a hundred years (Mair 2006), and we have the larger but less tidy general corpora (like the Bank of English) that use a wide variety of texts available in electronic form and are less ambiguous in their attempt to provide a socio-linguistically sound stratification than the BROWN family. The big but messy type of corpus has gained a lot of attention through the availability of large amount of data that can be downloaded from the world-wide web. Each corpus type has its advantages and disadvantages. The small but tidy type allows us to

'watch' the diffusion of grammatical innovation through different text and user variables. The large but messy type allows us to follow less frequent phenomena in lexical as well as idiomatic meaning. In the near future we may have a chance to combine the advantages of both corpora types by expanding ICE in a historical dimension and supplementing the existing ICE corpora by larger monitor corpora from the world-wide web. This will allow us to start more comparative ICE studies than before, which offers new perspective in applied as well as theoretical issues.

References

Granger S. (ed.), 1998, *Learner English on Computer*, Addison Wesley Longman, London/New York.

Greenbaum S., 1988, "A proposal for an International Computerised Corpus of English", *World Englishes*, 7, 315.

Greenbaum S., 1996, *Oxford English Grammar*, Clarendon Press, Oxford.

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

Greebaum, S., Nelson G. (eds), 1996, *World Englishes*, 1, Special issue on *Studies on The International Corpus of English*.

Hudson Ettle D., Schmied J., 1999, *Manual to Accompany the East African Component of The International Corpus of English*. Chemnitz University of Technology, Chemnitz, http://www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/ICE-EA/index.html, accessed December 1, 2010.

Kachru B., 1990, *The Alchemy of English: The Spread, Functions, and Models of Non-native Englishes*, University of Illinois Press, Urbana-Champaign.

Mair C., 2006, *Twentieth-Century English: History, Variation and Standardization*, Cambridge University Press, Cambridge.

Mesthrie R., Bhatt R.M., 2008, *World Englishes. The Study of New Language Varieties*, Cambridge University Press, Cambridge.

Nelson G., (ed.), 2004, *World Englishes*, 23, 2, Special issue on *The International Corpus of English*.

Nelson G., 2010, "ICELight: An Internet-sourced International Corpus of English", *ICES – International Corpus of English Studies*, 1, 1, pp. 12-21.

Nelson, G., Wallis S.A., Aarts B., 2002, *Exploring Natural Language: Working with the British Component of the International Corpus of English*, John Benjamins, Amsterdam/Philadelphia.

OED: *Oxford English Dictionary*, http://oed.com/.

Quirk, R., Greenbaum S., Leech G., Svartvik J., 1985. *A Comprehensive Grammar of the English Syntax*, Longman, London.

Schmied J., 1989, "Compiling a Corpus of East African English", *ICAME*, 13, pp. 75-77.

Schmied J., 2010, "English as a Lingua Franca of the European Union? Debates, Concepts, Projects", in M. Albl-Mikasa, S. Braun, S. Kalina (eds), *Dimensionen der Zweitsprachenforschung – Dimensions of Second Language Research. Festschrift für Kurt Kohn*, Narr Francke Verlag, Tübingen, pp. 141-158.

Schneider E. W., 2004, "How to Trace Structural Nativization: Particle Verbs in World Englishes", *World Englishes*, 23, 2, pp. 227-249.

Schneider E.W., 2006, *Postcolonial English. Varieties around the World*, Cambridge University Press, New York.