World Englishes, Vol. 9, No. 3, pp. 255-268, 1990.
Printed in Great Britain.

0883-2919/90 $3.00 + 0.00

Corpus linguistics and non-native varieties of English

JOSEF SCHMIED*

ABSTRACT: This article derives from the internal discussions of a project that has just been launched and which may provide a useful example of modern comparative linguistics: the International Corpus of English (ICE). It concentrates on the problems which arise when the principles of corpus compilation, which were developed in native communities (ENL corpora) in the pre-sociolinguistic age, are applied to non-native communities (ESL corpora) such as Africa. In my opinion this reveals a crucial difficulty in corpus compilation that has been neglected in most corpus-linguistic work: the contrast and relationship between variation according to use and that according to user, or between stylistic sampling categories based on text types and sociolinguistic ones based on speaker/writer identity. Examples of such problems will be derived from the second-language corpus I am primarily concerned with, the Corpus of East African English, but the principles of socio-stylistic variation in native and non-native varieties of English go far beyond this immediate context. They aim at combining two modern quantitatively oriented linguistic subdisciplines to their mutual benefit. After a brief introduction to the ICE project the following points are dealt with: first, the uses of computer-readable corpora for modern grammars and dictionaries in general (Section 2) and for applied (Section 3) and theoretical (Section 4) research on non-native varieties of English in particular, then the text type approach applied in ENL corpora so far (Section 5) and the sociolinguistic dimension with its relationship to stylistic variation (Section 6), followed by practical considerations for Third World Englishes (Section 7), and finally a multidimensional approach to socio-stylistic variation (Section 8) which may be necessary for transferring the ENL-based methodology of corpus compilation to ESL varieties.

1. INTRODUCTION[1]

The International Corpus of English (ICE) was launched and is co-ordinated by Sidney Greenbaum [from University College, London; cf. Greenbaum (1988a)], who also edits an ICE Newsletter documenting the discussions and development of the project. Its basic idea is that during the next few years machine-readable corpora of English will be compiled for the UK, the USA, Jamaica, Canada, Australia, New Zealand, the Philippines, Hong Kong, Singapore, India, Nigeria, East Africa and possibly for other countries, which will be supplemented by a corpus of international English, i.e. of English used in communication between speakers/writers of more than one nation, a corpus of EFL texts and a corpus of translations into English. These corpora will be compiled according to the same principles of informant/text selection and processed according to the same principles of coding and machine-readable storing. The text collection and recordings are to be made between 1990 and 1993; the transcription, the automatic tagging with grammatical word categories and the international distribution of such a tagged corpus (through ICAME in Bergen) should be finished by about 1995. Each national corpus will be composed of 250 spoken and 250 written texts of approximately 2000 words and will thus amount to 1 million words [cf. Greenbaum (1990)].

2. MACHINE-READABLE CORPORA AS THE BASIS OF MODERN GRAMMARS AND DICTIONARIES

Before a corpus can be compiled its possible uses must be considered, because "the preparation of a corpus cannot be seen in isolation from its intended uses" (Johansson,

*English Linguistics Department, University of Bayreuth, 8580 Bayreuth, FRG.

1978: 1). A simple example to illustrate this point here is that the corpus size required is determined by the purpose: frequent phenomena necessitate less text material than infrequent ones and grammatical analyses usually less than lexical ones—1 million words is in most cases highly sufficient for the former, whereas some 20 million may not be enough for specialized research in the latter.

In general, intervariety analyses can be distinguished from intravariety ones. So far, the comparative analysis of several varieties has been fairly limited. Work has been concentrated on the American Brown University Corpus (BROWN) and the British Lancaster–Oslo/Bergen Corpus (LOB); the analysis of the Indian Kolhapur Corpus has only just begun [cf. Shastri (1988)]. Intravariety analyses are much more frequent and further developed; they can therefore be used to show some practical realities of corpus-based work on (British) Standard English before an outline is given of some future possibilities of work on African English.

Machine-readable text corpora form the basis of all large-scale modern lexicographical and grammatical work on solid empirical foundations. Some major recent publications about "Standard English" (the British and American ENL varieties) from such a corpus-linguistic perspective illustrate this.

The *COBUILD English Language Dictionary* published by Collins in 1987 is only one of the publications that have appeared as a result of the joint efforts of Collins and a team of researchers led by John Sinclair at Birmingham. Their aim was to build up a database of 'real language' as it is called in the advertisements, so that grammar and dictionaries could be "compiled by the thorough examination of a representative group of English texts", allowing the researcher direct "daily access to 20 million words, with many more in specialized stores" (p. xi). The corpus-linguistic basis of such a dictionary facilitates the choice of authentic sample sentences and of lexemes according to frequency statistics—if the corpus text selection is broad enough to be representative. It almost goes without saying that the competing *Longman Dictionary of Contemporary English* is based on a corpus of similar size, the *Longman Citation Index.* Intended more as a dictionary for foreign learners than the first its strength lies in a further corpus-linguistic application, that is, its computer-controlled use of defining vocabulary checked with reference to a limited lexicon of the most frequent and necessary words for this purpose.

The *Comprehensive Grammar of the English Language* (CGEL) [by Quirk *et al.* (1985)] published by Longman also has a corpus basis, as the cover emphasizes: "The grammatically analysed data in the Survey of English Usage, together with other collections of material, both American and British [the BROWN and the LOB corpora mentioned], has again proved an invaluable source of new insights, and, equally important, has provided authentic examples of use." Therefore the CGEL promises not only "up-to-date and authentic treatment", but also "more grammar than meets the eye", which means that it is possible to calculate from a database quantitative differences in the occurrence and co-occurrence of syntactic and semantic features invisible to the linguist's eye or untraceable for the linguist's introspection.

The CGEL also compares some frequency counts in the American BROWN and the British LOB corpus, which can serve as an illustration of the basic idea of an intervariety comparison. Interestingly, differences between British and American English (BrE and AmE) are not mentioned in the main text but only in notes; they can be found in all word classes, such as verbs, pronouns and prepositions: among the modals, *shall/should* occur much less frequently in British than in American English (p. 3.39n); there is a much stronger

tendency towards *-one* in BrE than in AmE where forms with the rival *-body* have almost the same frequency in some compounds (p. 6.46n); prepositions usually have the same frequency in both varieties [for example *under, below, beneath, underneath* (p. 6.19n)], except for the Briticisms *amongst* [instead of *between, among,* and other (p. 9.21n)] and *round* (as opposed to *around*). Although many more differences in lexeme choice and grammatical structures could be expected the CGEL has used these possibilities only to a limited extent, because it concentrates on the common core (p. 1.42). This indicates the possible direction of future expansion.

Whereas grammars usually emphasize the common features that unite the different 'Englishes' and will thus have few structures marked 'BrE' or 'AmE' and only a few more 'esp. BrE' or 'esp. AmE', one of the principal purposes of dictionaries has always been to differentiate between the general lexicon and those entries that are specific to certain styles, registers, regional and social groups, etc. The biggest new lexicographical project that could include all this is the *New Oxford English Dictionary* (*New OED*) [cf. Johansson (1988)]. It also employs, in its second phase, a modern electronic database including a database management system, which will allow the dictionary user, for instance, to search specifically, even on-line, for entries with a stylistic/register or regional/social label. Finally the *New OED* could even be supplemented one day by 'satellites', i.e. dictionaries of regional varieties. The ways in which such a dictionary can be expanded, updated, and applied opens up a new world of lexicography.

From a sociolinguistic perspective most corpus-linguistic analyses and applications have been based on linguistic variation within a given corpus classified according to text types of a relatively homogeneous standard language. This was hoped to be ensured by the collection of texts from educated speakers only, whereas the variational aspects were thought to be adequately covered by the text types set. The few studies that have explored intervariety differences, mainly between the Standard varieties of AmE and BrE, treated Standard English as one sociolect of educated English. This does not mean that variation was neglected entirely, because "the corpus linguist should also deal with those (pragmatic) aspects of the situation that are linguistically relevant" (Aarts and van den Heuvel, 1985: 306), but the question arises as to which extralinguistic, variational aspects can be considered linguistically relevant. In corpus-linguistic work so far this has mainly been seen in terms of register, medium and style.

3. CORPORA OF NON-NATIVE VARIETIES OF ENGLISH AS A BASIS FOR THEIR PRACTICAL GRAMMATICAL AND LEXICOGRAPHICAL DESCRIPTION[2]

The examples mentioned illustrate that recent dictionaries and grammars of current Standard English (ENL varieties) are based on modern corpora. As far as ESL varieties are concerned, studies leading to the compilation of similar inclusive and exclusive dictionaries [cf. Görlach (1990)] and grammars would be feasible if the necessary corpora were available and if research teams and publishers were interested in corpus-linguistic work. But there are also smaller and more immediately practical corpus-linguistic projects possible that may be considered by individual ESL researchers.

It will, for instance, be possible to locate unusual words in a non-British/non-American corpus by applying orthographic programs (as are available in any good comprehensive text processing program today, such as *Word* or *Word Perfect*) which compare texts with Standard English dictionaries. This application could lead to the compilation of a complete

dictionary of regional English given the possibility of integrating enough texts into the corpus and running them through the program. Ideally, this could be accomplished by co-operation with, for instance, a modern newspaper that uses computers to produce its texts and layout in machine-readable form. The newspaper could provide the texts and would be offered the services of the new dictionary of regional English in return.

To make a more balanced and specific investigation into the lexicon of non-native Englishes, e.g. a certain semantic field such as European and African food terms, concordance programs (e.g. the Oxford Concordance Program) can be applied and the occurrences (or non-occurrences) of particular lexical items determined, not only their frequencies, but also their collocations and environments. Thus it is also possible to find out more about the internal structure of the lexicon, whether words of similar meaning behave syntactically differently or occur in different contexts, whether StE lexemes are expanded or restricted in non-native usage, in what way heteronyms [lexemes of equivalent meaning, but different social/geographical distribution; cf. Görlach (1989)] can be differentiated, and so on.

By applying more sophisticated statistical programs (e.g. *SPSS/PC+*)—possibly combined with the tagged version of a corpus and better analytical tools (e.g. of *Word Cruncher*) in the future—three types of analyses can be undertaken: semantico-syntactical, textual and sociolinguistic ones. Co-occurrences of certain semantic and grammatical features, the relative frequencies of (near-)synonyms, or even the density and consistency of 'nativeness' [i.e. variety-specific features, like Indianisms or Briticisms, cf. Algeo (1988)] of a particular text can be calculated. A comparison is possible between the occurrences of phenomena in more informative and more imaginative texts; the occurrence of formal features in informal texts can be examined; the development and frequency of features in texts of different social or educational levels may be followed, and so on.

In addition to these analyses of intravariety variation the real strength of the ICE lies, of course, in its possibilities for intervariety comparison. Besides the intervariety 'deviation' of non-native from native Englishes mentioned so far, intervariety comparisons between non-native Englishes may also be valuable as, for instance, when practical decisions about standards for a certain level of education must be taken. If African teachers have to decide which deviant structures to 'unteach' in specific exercises and which to accept, they may be inclined to accept features much more readily if the same ones occur in other Englishes, in Indian English, for example. In such an approach interference phenomena from the first language may be less tolerated than general learning strategies in New Englishes. If the national functions of English are emphasized common national features may be much more readily accepted than subnational ones associated with particular ethnic groups.

In general, corpus-based studies tend to be related to formal surface structure, although linguistic programs [like the LDB (Linguistic DataBase from Nijmegen)] will make work on sentence structure and functional patterns possible and thus facilitate comparisons in sentence/text complexity. But for the application of linguistic work in the African context the restriction to surface structure may not be a disadvantage, as this makes its direct relevance to practical grammatical and lexicographical problems more obvious.

4. THE IMPORTANCE OF CORPUS-LINGUISTIC WORK FOR RESEARCH ON NON-NATIVE ENGLISHES

From these practical examples of corpus application it becomes clear that corpus-linguistic analyses can also make a decisive contribution to research into the 'nativization' [e.g. Kachru

(1986)] of English in African or Asian sociocultural contexts. One obvious advantage is that they allow the researcher to confirm quantitatively and statistically impressions he has gained from introspection or participant observation about qualitative differences between native and non-native Englishes. But even below this openly perceptible level of usage, hidden, unconscious structures of language can be discovered. The subtle processes of nativization can be observed on the levels of relative frequency of lexemes and constructions or of collocations and so on.

As sociolinguistic analysis is quantitative, it must operate on a broad, more or less standardized empirical basis of texts compiled according to the linguistically relevant parameters of variation. For most analyses beyond the level of pronunciation a corpus, its size depending on the type of analysis envisaged, constitutes at least a good starting point—it may be complemented later by phenomenon-specific text selections. As the analysis of non-native Englishes has concentrated on the more salient features of pronunciation, loan words and idiomatic expressions, grammatical analysis is still underdeveloped. In this area it would be particularly useful to prove the general assumptions that non-native Englishes form clines of intelligibility and develop their own systematicity. Only on a corpus basis will it be possible to answer questions such as the following: to what extent has the definite–indefinite distinction of Standard English been replaced by a specific–non-specific distinction in non-native varieties (Platt *et al.*, 1984: 54)? How far have the progressive forms (*be* VERB-*ing*) really been expanded *(ibid.:* 72), for which verbs in which verb patterns and with which collocates? Then it may also be feasible to draw up a hierarchy of features of 'Africanization', 'Indianization', etc. and a hierarchy of texts exhibiting 'Africanness', 'Indianness', etc.

Such questions of systematicity can only be answered in a systematic quantitative framework such as a corpus. Then the research on non-native varieties of English which is presently dominated by qualitative aspects can be complemented by quantitive ones. Only a quantitive comparison of stylistic features of native and non-native varieties will, for instance, show to what extent non-native varieties have really the restricted stylistic variation that has often been claimed or whether all text types have the same grammatical characteristics in native and non-native varieties.

These quantitative measurements of variation in performance can then also be applied to draw up guidelines for the new norms to be institutionalized. For this it would also be important to confirm that an association between judgments of frequency and judgments of acceptability exists, not only in native (Greenbaum, 1988b: 94–112) but also in non-native varieties.

Finally, another aspect of corpus-linguistic research in non-native varieties may be worth mentioning, which is perhaps, at the moment, only an envisaged goal. If corpus- and sociolinguistic research can really advance at a coordinate pace a sociolinguistic text basis may one day be used to extract information about an unknown writer purely on the basis of formal textual criteria.

Before, however, such fruits can be harvested from the ICE it has to be compiled. The following section will therefore discuss some theoretical problems[3] related to the basic compilation principles of the corpus, which are concerned with internal variation within a second-language variety. It will not only question whether the suggested corpus structure is internally conclusive and logical, but also whether the linguistically relevant parameters in second-language societies differ from those in first-language societies.

5. THE TEXT TYPES USED FOR MACHINE-READABLE CORPORA[4]

The first corpus of English, the Survey of English Usage (SEU), was started by Randolph Quirk as early as 30 years ago—although the collection of oral and written material and even more so its computerization has only recently been completed. The spoken SEU texts are subdivided into 'dialogue'–'monologue' and the written ones into 'written for spoken delivery'–'non-printed'–'printed'. More restricted principles of text collection were applied to the BROWN and LOB corpora, for American and British English, respectively; both are only concerned with written language and consist of texts from only one year (1961). The Kolhapur Corpus, which compiled texts from 1978, had to accept the BROWN–LOB categories in order to remain compatible, although some text types were difficult to find in India and the number of texts in some categories had to be changed (Shastri, 1988)—early evidence of the difficulties related to an ESL corpus that this article focuses on.

The text types of the BROWN–LOB–Kolhapur corpora do not cover all areas represented in the SEU, comprising neither spoken nor non-printed language. They are pragmatically subdivided by several overlapping parameters: printing type (newspapers–books–government and other publications), readership (popular lore–scientific writings) and topic/content area (religious–science fiction–humour).

Because the SEU approach covers a wider area of language variation the basic typographical categories of the ICE will be derived from the SEU along the lines 'spoken'–'written','monologue'–'dialogue', 'printed'–'non-printed'. This basic structure is fairly uncontroversial, as there is enough empirical evidence that these categories have at least some characteristic formal consequences in terms of linguistic features.

The importance of the distinction between spoken and written language is based on the following parameters:

> EXTRALINGUISTIC SITUATION, i.e. in the oral medium speaker and listeners are aware of their immediate environment and do not have to refer to it explicitly; in the written medium the use of more explicit adverbials of time or place is necessary;
> TIME, i.e. in the written mode the writer can plan his sequence of contents and his linguistic means more carefully, while the reader can choose his own speed of decoding the message, go back to previous passages, compare, skip, stop and think, etc. (this makes more subordination possible, for instance);
> FEED-BACK, i.e. questions and criticism are easily possible in spoken dialogue, but troublesome in written, that is why possible criticism is anticipated and possible misunderstandings are taken into consideration in planning the written form (this implies, for instance, the choice of words).

Although many paralinguistic signals (in speed and voice quality) and visual signals (like mime and gesture) are lost in the written medium, some features of spoken language have more or less equivalent written counterparts. Intonation can, for instance, be partly rendered by punctuation and syntactic means. Specific features of spoken and written English will be found, though not on all linguistic levels to the same extent, in the respective ICE texts and both kinds or intermediate features in the 'intermediate' text types 'spoken for written delivery'[5] and 'written for spoken delivery' (or 'scripted'), be it on radio, TV, film or stage. Written language is usually associated with the parameters PREPARED, INDIRECT, PUBLIC, FORMAL, spoken language with SPONTANEOUS, DIRECT, PRIVATE, INFORMAL.

The distinction between dialogue and monologue is considered crucial because in a dialogue linguistic features can be taken up from the communication partner and may thus be different from (at least in this aspect) uninfluenced monologue. The distinction between non-printed and printed is crucial because it can be assumed that much more attention is paid to printed than non-printed language; a writer's draft may even be checked and changed by others. At the level of subcategories the system soon becomes more complex.

The general parameters behind such subcategorizations are expressed more or less explicitly, but usually include the variables:

public–private (the main distinction in spoken dialogue/monologue and scripted texts) refers to the presence or absence of (many) other people apart from the speakers participating in the dialogue and that the speakers are aware of that and (re-)act accordingly, i.e. monitor their language more closely;
direct–distanced (the subordinate distinction in spoken private monologues and dialogues) draws attention to the possibility of using technical means such as the telephone (including conference calls) and the tape-recorder; here the situational context and some of the listeners' reactions are lost, but through the telephone both partners can still communicate; the listeners cannot communicate with the speakers when the tape recorder or the broadcast system (under dialogue spoken public) is used, where the listeners are also invisible to the speaker(s);
spontaneous–prepared (used to subcategorize spoken public monologues) is related to the important linguistic hypothesis that the more attention is paid to speech the more formal it will be;
business–social (used to subcategorize letters) is a heterogeneous label as it includes aspects of the distinction public–private, of the social distance between communication partners and of the different topics mentioned;
informational/persuasive/administrative/instructional/imaginative (the main subcategorization for the printed texts) refers to the purpose of the text, which is also related to the content and formality of style;
learned–popular (used as part of the subcategorization of informational texts) refers to readership, with important consequences in formal style.

This interpretation of the suggested ICE subcategorization is by no means exhaustive, but it shows the main principles. These principles can be used to cover as wide a range of texts as possible and to characterize these texts according to their 'style' as a complex combination of parameters, which can be supplemented by others related to the participants involved in communication.

6. THE SOCIOLINGUISTIC EMPHASIS AND SOCIO-STYLISTIC VARIATION

In contrast to the discussed categorization according to language uses microsociolinguistic variation is usually seen along the lines of sex, age, education/socioeconomic status, and region/ethnic affiliation of language users. These categories have been established in sociolinguistic research in ENL societies and were more or less confirmed by research in ESL societies, although their relative importance and necessary additional variables may be a matter of debate.[6] That is why it may be safe for the ICE to concentrate on them and to select informants along these dimensions, too. A grid of these dimensions can be laid over a 'second-language society' to capture the variation in their non-native English usage, which should, of course, be reflected in the compilation strategy of the corpus texts. These sociolinguistic parameters thus have to supplement the text type parameters discussed above.

In a much wider sociolinguistic framework, as offered by Dell Hymes's approach, the ethnography of speaking, it should easily be possible to combine the (narrow) sociolinguistic with the text type categories. This comprehensive approach would include:

participants: speaker, hearer(s), listener(s); their features in terms of sex, age, education, languages spoken, etc.; and their relationship;
form of communication: medium (spoken, written), channel/means (direct, distanced);
speech-act: topic (e.g. religion, business), aim (e.g. instructional, informational), function (e.g. information, action, maintenance of social relationship).

In this way the usual sociolinguistic investigations, with emphasis on Dell Hymes's categories under 'participants', and the suggested text type approach, with elements from Dell Hymes's 'form' and 'speech-act' categories, can complement each other. In a comprehensive sociolinguistic approach all these parameters have to be considered.

The relationship between stylistic and social variation has been represented in a Labovian paradigm as parallel clines from most informal casual speaking to most formal minimal

pair reading style on the one hand and the social continuum of the informants on the other. Since Labov's study it has been argued that speaking and reading may constitute two separate continua as the underlying factor assumed 'the amount of attention paid to speech'. As an alternative suggestion, 'audience design' has been put forward (Bell, 1984), which sees variation in speech mainly as a response to other people, along with other kinds of interactive behaviour. Despite this criticism of some Labovian assumptions, however, the basic quantitative techniques of the Labovian approach should not be rejected, because they permit a comparison between the behaviour of an individual on different occasions with that of different individuals on a single occasion, i.e. the interrelationship between interpersonal and intrapersonal variation [cf. Milroy (1987: 172–183)]. A very general sociolinguistic result has been a shift towards standard forms in more 'formal' contexts, which may be caused by the participants, the topic or the setting. This shows that the concept of formality is a complex one and that its influence may be manifold, and even indirect and varied on the different linguistic levels.

In a non-native context the problems of learner languages and code-switching arise: In what way are the stages of the learner continuum from acrolang to basilang related to the formal–informal cline? And if they are, can speakers with less education manage a more formal language if such is demanded by the topic or situation, and do educated speakers 'relapse' into more mesolang varieties when conversing with less educated people or are those not used at all because then African languages are used instead of English?

As the sociolinguistic principles have been worked out mainly on the basis of investigations concerning pronunciation in Western societies and/or among first-language speakers [cf. the contributions in Romaine (1982) or Cheshire (1990)] it would obviously be an important part of a second-language corpus to test some of them in grammatical and lexical analyses in Third World societies [cf. Schmied (1989b)]. This is a strong argument in favour of including non-native corpora in the ICE and for including as many sociolinguistic criteria as possible.

7. PRACTICAL CONSIDERATIONS FOR CORPORA OF NON-NATIVE ENGLISHES[7]

One solution to these partly conflicting principles in corpus compilation seems to be a compromise: whereas broad text type categories are used for the larger categories the subdivision could be made according to broad sociolinguistic principles. But then sampling becomes much more difficult because, for instance, a business letter written by a speaker of a particular first language with secondary school education in his early twenties would have to be found. This example shows that such a compromise cannot follow the sampling principles too rigidly, but an attempt at including sociolinguistic text characteristics consciously among the compilation principles for a second-language corpus would be of value. In the non-native context the strict sociolinguistic criteria of a random stratified sample, where the percentage of texts would correspond with the percentage of the population, cannot be followed but a deliberate attempt can be made to achieve a so-called 'quota stratified sample', which would aim at including different values for a certain variable consistently. This means that no more texts with a particular value are taken when this value occurs already a certain number of times; texts with other values for this parameter are chosen instead. If, for example, a large number of the text samples for student essays comes from a particular language group, only texts from other groups will be added in this category to avoid overrepresentation. In other text categories it may be almost

impossible to control the variables tightly. In general, non-native corpora of English should include the major language/ethnic groups and some age groups other than the easily accessible student population.

On the other hand, it may not be necessary to follow all the text types rigidly either, because at least some text types may internally be more homogeneous than others. It is, for instance, difficult to determine all the variables that play a role in public dialogue discussions. Some linguists have even pointed out that the introductory and final passages of certain texts have their own style and have argued in favour of including only central passages. As this cannot be done in many cases the position of the excerpt in the complete text should be indicated in the text description, so that it could be used as a variable in a specific analysis.

It is important to note that some of the subcategories used for ENL corpora are also based on important sociolinguistic distinctions, such as the distinction between face-to-face conversations (suggested in other versions of ICE text type proposals) among equals (colleagues, friends etc.) and those among disparates (professional–client, parent–child etc.), and these certainly have to be marked carefully in the ESL corpora, whereas others such as the distinction between sermons and lawyers' presentations in court seem to be less important for a general corpus and related to specific professional registers. A corpus that is relatively small will certainly not be sufficient for investigations into professional language, and it will only become clear from investigations into larger corpora whether this text type really has specific syntactic features. A restriction to more general text types may therefore be justified not only from the practical but also from the theoretical side. Some distinctions such as that between private dialogues over the telephone and direct ones are doubtful, because they may simply not play a role in Third World societies (and may thus not have the same characteristic stylistic features), or they may even be impossible, from a practical point of view. Phone-in programmes may be an interesting cultural phenomenon, but again it is not clear whether it is really worth an attempt at including them because they would be hopelessly overrepresented in the Third World context.

As was indicated above, in order to decide on a viable compromise between all the parameters suggested it is necessary to ask: What is the collected corpus for? In what way is it to be used for analysis later? As Romaine (1982: 112f) has argued, there is a scale of linguistic features ranging from those that are fairly independent of style, e.g. on the level of pronunciation, to those that are to a great extent style-conditioned, e.g. vocabulary or sentence length. Interestingly enough, the cline of sociolinguistic relevance seems to run in the opposite direction: pronunciation scores high and vocabulary and sentence length low. As the planned corpus will not be big enough for specialized vocabulary analyses it may not be of advantage to subdivide the categories too much according to stylistic dimensions. If the suspicion arises that a certain parameter may have important formal repercussions and not enough tokens are available, expanding the corpus in precisely that direction for specialized analysis may be expedient.

One way of providing a basis for this is to encompass the national ICE corpora with possible additional texts in a larger monitor corpus, which can be consulted for certain purposes if results from the actual corpus suggest the need to do so. As some linguistic features occur only in certain text types there must always be enough texts of all types. That is why the ICE plans to have at least 10 texts in each category. When the social parameters are included the problem is aggravated. It may be possible to analyse the relative proportion of certain structures in *all* written texts (e.g. initial or final positions of

prepositions in relative clauses), but impossible to analyse the distribution of a particular feature across the speech of male/female speakers with different educational and first-language backgrounds. Finding enough tokens in the many cells in the sex–education–first-language grid may be problematic. Although this would give us valuable insights into the development of the feature, such detailed studies cannot be based on the corpus texts alone. Including texts from the larger monitor corpus can generally be considered fruitful for several reasons, either if the number of texts in some cells seems too small when certain parameters are investigated, or when some texts occur that seem to be relevant in the specific language community, but cannot be integrated into the general overall system.

The main problem with the development of culture-specific text types is the well-known corpus compilation paradox that a representative corpus for an analysis of the (socio-)linguistic structure of a variety can only be obtained if the compilation accords with the variety's main (socio-)linguistic principles. These principles and their relative importance are, however, precisely what linguists want to establish by analysing a representative corpus. Extremely formulated, that would mean that the second corpus for a variety will be much better than the first, the latter being only used to establish the main parameters for the next one. 'Theoretical sampling' can only be started once the variables are known. That means a qualitative analysis is required to establish the parameters first before a comprehensive and representative quantitative analysis can be attempted.

8. A MULTIDIMENSIONAL APPROACH FOR TEXT CATEGORIZATION

As there is no satisfactory comprehensive theory of socio-stylistic variation for speech and writing available and as it has been shown that text type categories can be broken down and defined in several parameters, a multidimensional approach has to be adopted. This means that all possible influencing parameters have to be recorded carefully, irrespective of whether they are thought to be axiomatic or not. The following list shows which parameters could be used and how they could be used to characterize text types compiled:

SPOKEN refers to the spoken medium;
2SPEAK implies the presence of other speaker(s) (as in the dialogue–monologue contrast);
LISTEN marks the presence of non-speaking listeners (as in private–public);
WRITTEN again refers to the medium (with the implications discussed above);
PRINTED usually implies more careful language, because it is proofread by the writer or editor;
DISPHYS refers to the physical distance between speaker(s) and listener(s) (as in broadcasts and with printed materials); this usually entails a clear and explicit form since the message has to be understood without the contextual support of the situation;
AUDSPEC means that the text is directed towards a specific audience/readership, which affects the degree of technicality adopted (this includes the learned–popular distinction above);
PLANNED (or prepared) refers to the amount of time permitted for text production (as in the spontaneous–prepared dichotomy above);
DISSOC records the social distance between speaker(s) and listener(s); it largely determines the degree of formality adopted (this includes partly the business–social distinction above);
TOPIC relates to a particular subject matter on the humanities–social sciences–natural sciences–technology scale;
INFORM is related to the dimension of interactional (sociolinguistic) and informative functions of language, e.g. greetings serve only an interactional and administrative forms mainly an informative purpose; the INFORM scale ranges from complete message orientation to complete hearer orientation and includes the differences between administrative–instructional–persuasive–imaginative texts;
CREATIVE refers to the stimulation or restriction of formal creativity and is related to the personal subjective or objective treatment of a topic; interestingly, both ends of the INFORM scale restrict creativity, either because the message must be delivered as plainly as possible or because social rituals make the inclusion of personal matters impossible.

These parameters can be found in various degrees in the following text categories (with main examples in parentheses):

A = spoken dialogues private
B = spoken dialogues public (e.g. broadcasts)
C = spoken dialogues public instructional (e.g. class lessons)
D = spoken monologue private
E = spoken monologue public (e.g. speeches, lectures)
F = written for spoken dialogue (e.g. drama)
G = written for spoken monologue
H = written non-printed public (e.g. student essays)
I = written non-printed private (e.g. letters, notes)
J = printed informational learned (e.g. M.A. theses?)
K = printed informational popular (e.g. school-books)
L = printed informational reportage (e.g. news)
M = printed documentary writing (e.g. legal/administrative reports)
N = printed creative writing (e.g. novels, stories in magazines).

The combination of these parameters and text types provides a grid (cf. Table 1) which is not only a reformulation of the ICE types but also offers suggestions of possible parallels and contrasts between the occurrences of particular features in the texts.

In addition to the textural variables the social features can be spread over the corpus texts. But due to their textual nature it is not possible to collect certain features for some texts. If the speakers or writers cannot be questioned about social criteria, which is usually the case in all texts with the parameters PRINTED or DISPHYS (e.g. from television), information about the social parameters can only be derived from their names. These indicate sex, where broadcast pictures and voice quality can also be used as indicators, and in most East African cases also first language (ethnic affiliations) or at least the dominant language background if an unusual name was adopted. More problematic is education, which can only be deduced very indirectly from occupations or special social positions. In the case of written non-printed texts social features may possibly be included, either derived from information given by the writers themselves, by those who provided the texts (editors, publishers, friends etc.) or they can be identified because they are publicly known, as is the case with famous writers and personalities.

Although this multidimensional categorization, including its distinctions and clines, has to be confirmed empirically and statistically, it is a possible starting point for the corpus-linguistic analysis of language variation that takes text types as a basis for the analysis of linguistic features.

This view is supported by empirical research that chooses the opposite direction, concentrating on grammatical features and using them to establish groups of text with common characteristics. The first steps in this direction were taken in several publications by Biber (e.g. 1986, 1988 or 1989) in a multi-feature/multi-dimension approach to linguistic complexity [cf. also Preston (1989: 265–269)]. He uses "factor analysis for empirical identification of groups of linguistic features which occur in a high frequency in texts, indicating a communicative function shared by these features" (Biber, 1986: 385). There is no need to go into the empirical details here; suffice it to say that he suggested three types of textual dimension: interactive vs edited, abstract vs situated and reported vs immediate. Some of his parameters were 'general audience' (my category +AUDSPEC), 'shared background' (contextual and cultural, the former expressed in my category +LISTEN, the latter suppressed, because it endangers crosscultural comparability), 'spontaneous vs planned' speeches (my category +PLANNED), 'interactional vs informational' emphasis (my category +INFORM).

Although the multidimensional approach offered here because of its desirability for ESL

Table 1. Textual and social variables and their application to the suggested text categories*

	Text													
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
	Number of texts													
Variable	100	40	40	20	50	10	40	40	20	40	40	20	20	20
SPOKEN	+	+	+	+	+	+	+	−	−	−	−	−	−	−
2SPEAK	+	+	+	−	−	+	−	−	−	−	−	−	−	−
LISTEN	−	+	+	−	+	+								
WRITTEN	−	−	−	−	−	+	+	+	+	+	+	+	+	+
PRINTED	−	−	−	−	−	−	−	−	−	+	+	+	+	+
DISPHYS	−	+	−	−	−	−	−	−	+	+	+	+	+	+
AUDSPEC	( )	( )	+	( )	( )	( )	( )	( )	( )	+	−	−	+	−
PLANNED	( )	(+)	(+)	( )	(+)	(+)	( )	(+)	(+)	(+)	(+)	(+)	(+)	(+)
DISSOC	(−)	(+)	+	(−)	(+)	( )	(−)	+	(−)	(+)	+	( )	(−)	( )
TOPIC	(#)	(#)	#	(#)	(#)	(#)	(#)	#	(#)	#	#	(#)	(#)	(#)
INFORM	(+)	(+)	+	(+)	(+)	(+)	+	−	(+)	+	+	+	+	(−)
CREATIVE	(+)	(+)	(+)	(+)	(+)	+	(+)	(+)	(+)	(−)	(+)	−	−	+
SEX	+	+	+	+	+	+	+	+	+	[+]	[+]	[+]	[+]	[+]
ETHNIC	#	#	#	#	#	[#]	[#]	#	[#]	[#]	[#]	[#]	[#]	[#]
EDUCATION	#	#	#	#	#	[#]	[#]	#	[#]	[#]				[#]
REGION	#	#	#	#	#	[#]	[#]	#	[#]	[#]				[#]
AGE	#	#	#	#	#	[#]	[#]	#	[#]	[#]				[#]
COUNTRY	+	+	+	+	+	+	+	+	+	+	+	+	+	+

*+/− = obligatory value used systematically to distinguish between text types; # = optional values in subcategories within text types used in stratifying variation within the text type; ( ) = indicated but not used to distinguish categorically, sometimes with a tendency towards +/−; and [ ] = indicated wherever possible.
NB: It goes without saying that wherever we have several speakers, as in the dialogue categories (A, B, C and F) the sociolinguistic features of all communication partners have to be included.

corpora differs—and has to differ—from the ENL-based SEU approach, the differences in compilation practice must be minimal, if the general goal of comparability is to be maintained. The multidimensional categorization may not only be more comprehensive but also more challenging than the BROWN/LOB text types. It includes the well-known SEU text types, but adds detailed text and sociolinguistic descriptions. Such a matrix approach has the additional advantage that it does not imply a hierarchy, e.g. that the written–spoken dichotomy is more 'basic' than the parameters of age or sex.

In view of the corpus compilation paradox mentioned above the first corpus of a language variety cannot be perfect, especially if the sociolinguistic principles of an ESL variety are also to be taken into account. But a satisfactory basis for many linguistic and sociolinguistic analyses can still be obtained. If desired, this may one day be expanded by specialized corpora or even converted into an open-ended relational database. The cross-references indicated, parallels and interrelationships between parameters, can also serve as hypotheses for empirical studies of various kinds.

In this short contribution I have tried to illustrate some of the problems, but also the fascinating possibilities, which could derive from corpus-linguistic research into non-native varieties of English and their comparison with each other or with native varieties. I hope to have shown:

—that compiling non-native corpora can be a challenging research goal;

—that we do not need to wait for a new theory of socio-stylistic variation, but can use the basic text type and sociolinguistic categories; they are, with minor

modifications, useful for the ESL context, even if we assume that the variation of educated non-native English occurs more along dimensions of language user than of language use; and

—that a comprehensive multidimensional approach could be used to allow various interpretations in future analyses.

NOTES

1. I am grateful to Professor Sidney Greenbaum for his various suggestions while I was writing this article. The following considerations are meant as a contribution from a sociolinguistic perspective towards the discussion of the sampling principles which the ICE is going to follow in the respective national (sub-)corpora. 'Corpus' in this narrow context means more than a collection of text material as a basis for linguistic analysis; it means that the compilation aims at a representative sample of language use within certain categories and that the texts are accessible in machine-readable form so that computer-based analytical tools can be applied. The contrast between variation according to use and according to user (Halliday *et al.*, 1964) has been described in various technical terms; for instance as diaphasic vs diastratic in Coseriu's terminology or diatypic vs dialectal in Gregory's. The categories of use are often called styles or levels or 'the five clocks' (Joos), codes (Gumperz) or register (Strang); they have been subdivided into such areas as field of discourse (topic), mode of discourse (spoken–written) and style of discourse (colloquial–primary) by Halliday *et al.* (1964). The CGEL (1985: 1.19) distinguishes region and social group from field of discourse (e.g. literary, religious, learned, scientific, instructional, technical, journalistic, bureaucratic; see p. 1.28), medium (spoken, written) and attitude (from formal to informal), and mentions varieties according to interference (from another language) separately (on p. 1.34).
2. For some linguistic phenomena corpus-based analyses are, of course, not particularly useful, either because their occurrence is so rare that even a linguist looking for a particular feature in his daily reading of books and newspapers will only come across it once a month or because the linguist is concerned with interesting semantic usages, or deep-structure phenomena, which escape the usual formal surface-oriented tools corpus linguists work with.
3. I wrote about the expected more practical problems of data collection when presenting the project of a Corpus of East African English (CEAE) for the first time in Schmied (1989a). As the CEAE will be integrated into the larger ICE a few technical changes in the categories and processing strategies have to be made, but the central problem, to what extent a national corpus should be kept compatible and to what extent it should be variety-specific, remains the same.
4. As far as the corpus approaches and text types are concerned, a similar survey, from a grammatical perspective, can be found in Oostdijk (1988).
5. In the ICE discussion the category 'spoken for written delivery', i.e. dictation, was taken as the (only) text type for the category 'private monologue', with the subcategories 'face-to-face' and 'into tape-recorder', but without the distinction used for letters (in the category 'written non-printed' between social and business letters. These text types are [after Crystal and Davy (1969: 66f)] complex according to the categorization by 'medium' which overlaps to a large extent with the monologue–dialogue distinction in '(speaker) participation' as oral texts tend to be dialogues, whereas written texts tend to be monologues.
6. It will, of course, be one of the main theoretical sociolinguistic uses of the non-native language corpora to investigate whether the sociolinguistic dimensions have the same relevance as outside native language corpora or outside the Western sociocultural context, e.g. whether socioeconomic variables are also prominent or whether ethnic ones are decisive, whether in contrast to most European studies female speech is really further away from the standard than male speech [cf. Russell (1982)] and so on.
7. This section will not deal with country-specific compilation problems, but rather with the general issue. We must also bear in mind that economic and sociocultural problems may play a decisive role. They may, for instance, affect the availability of certain text types. The number of books produced in Tansania is often so low that it may not leave much choice as to what can be compiled. Under these circumstances including excerpts from MA theses into the 'printed' informational learned text category may have to be considered. Cultural differences tend to be neglected in an international approach, but culture-specific text types (e.g. matrimonial advertisements in India) may valuably be included in the larger monitor corpus. Even the text categories may not have the same weight in non-native environments; as English is clearly related to the educational sector texts from related categories are more important than in ENL corpora.

REFERENCES

Aarts, Jan and van der Heuvel, Theo (1985) Computational tools for the syntactic analysis of corpora. *Linguistics*, **23**, 303–335.

Algeo, John (1988) A computer corpus of a Dictionary of Briticisms. In *Corpus Linguistics, Hard and Soft: Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora.* Edited by Merja Kytö, Ossi Ihalainen and Matti Rissanen. Amsterdam: Rodopi. pp. 45–59.

Bell, A. (1984) Language style as audience design. *Language in Society,* **13,** 145–204.

Biber, Douglas (1988). Spoken and written dimensions of English: resolving the contradictory findings. *Language,* **62,** 384–414.

Biber, Douglas (1988) *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, Douglas (1989) A typology of English texts. *Linguistics,* **27,** 3–47.

Cheshire, Jenny, ed. (1990) *English around the World: the Social Contexts.* Cambridge: Cambridge University Press.

COBUILD (COBUILD English Language Dictionary) (1987). London: Collins.

Crystal, David and Davy, David (1969) *Investigating English Style.* London: Longman.

Görlach, Manfred (1990) Lexical problems of English in Africa. In *Linguistics in the Service of Africa with Particular Reference to Research on English and African Languages.* Bayreuth African Studies Series 18. Edited by Josef Schmied. pp. 27–46.

Greenbaum, Sidney (1988a) Proposal for an international computerized corpus of English. *World Englishes,* **7,** 315.

Greenbaum, Sidney (1990) Standard English and the international corpus of English. *World Englishes,* **9,** 79–83.

Gregory, M. (1967) Aspects of varieties differentiation. *Journal of Linguistics,* **3,** 177–198.

Halliday, M. A. K., McIntosh, A. and Strevens, P. D. (1964) *The Linguistic Sciences and Language Teaching.* London: Longman.

Johansson, Stig (1978) Manual of information to accompany the Lancaster–Oslo/Bergen Corpus of British English, for use with digital computers. Oslo.

Johansson, Stig (1988) The *New Oxford English Dictionary* project: a presentation. *ICAME Journal,* **12,** 37–41.

Kachru, Braj B. (1986) *The Alchemy of English: the Spread, Functions and Models of Non-native Englishes.* Oxford: Pergamon Press. (Reprint 1990, University of Illinois Press, Urbana, IL.)

Milroy, Lesley (1987) *Observing and Analysing Natural Language.* Language in Society 12. Oxford: Basil Blackwell.

Oostdijk, N. (1988) A corpus linguistic approach to linguistic variation. *Literary and Linguistic Computing,* **3,** 12–25.

Platt, John, Weber, Heidi and Ho, Mian Lian (1984) *The New Englishes.* London: Routledge & Kegan Paul.

Preston, Dennis R. (1989) *Sociolinguistics and Second Language Acquisition.* Oxford: Basil Blackwell.

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

Romaine, Suzanne (1982) *Socio-historical Linguistics. Its Status and Methodology.* Cambridge: Cambridge University Press.

Romaine, Suzanne, ed. (1982) *Sociolinguistic Variation in Spech Communities.* London: Edward Arnold.

Russell, Joan (1982) Networks and sociolinguistic variation in an African urban setting. In Romaine (1982: 125–140).

Schmied, Josef (1989a) Conference report: compiling a corpus of East African English. *ICAME Journal,* **13,** 75–77.

Schmied, Josef (1989b) English in East Africa: theoretical, methodological and practical issues. In *English in East and Central Africa* I. Bayreuth African Studies Series 15. Edited by Josef Schmied. pp. 7–37.

Shastri, S. V. (1988) The Kolhapur Corpus and work done on its basis so far. *ICAME Journal,* **12,** 15–77.

Weiner, Edmund (1986) *The New Oxford English Dictionary* and world English. *English World-Wide,* **7,** 259–266.

(*Received* 24 *January* 1990.)