Compiling the Sri Lankan component of ICE: Principles, problems, prospects

Joybrato Mukherjee, Marco Schilk and Tobias Bernaisch, Justus Liebig University, Giessen

1 Introduction: ICE-SL reloaded

Ten years after the ICE project had been initiated by Sidney Greenbaum in the late 1980s (cf. Greenbaum 1996a), the idea of compiling a Sri Lankan component of ICE was also born. In the late 1990s, the first team began to collect data for ICE-Sri Lanka (ICE-SL) under the auspices of Christopher Tribble (pers. comm.), who at that time was based in Colombo. For various reasons, data collection came to a halt soon afterwards, and, in 2002, Tribble left Sri Lanka for Poland. In early 2005, when he heard about the research focus of English linguistics in Giessen on South Asian varieties of English, he suggested that the Departments of English of the University of Giessen and the University of Colombo take over and resume the compilation of the corpus – and, in fact, in 2006 ICE-SL was reloaded as a joint project of the two universities with some initial funding provided by the University of Giessen for the written part of ICE-SL.

In the present paper, we will discuss various aspects that have been relevant to – and have had implications for – the compilation of ICE-SL since 2006. We will start off from a general overview of the status and role of English in present-day Sri Lanka, including some comments on the changing language policy in the post-independence period, in order to illustrate why English in Sri Lanka ought to be covered by ICE (Section 2). We will then address some general principles and problems that the project team had to face when compiling and annotating the corpus, e.g. the question of what kind of English and whose English should be represented and the issue of how data for individual genres can be gathered (Section 3). Afterwards, we will sketch out the current state of the written part of the corpus and provide an overview of descriptive studies that have already been conducted on the basis of pilot versions of ICE-SL (Section 4). We will then discuss some of the methodological issues that will be relevant to the compilation of the spoken part of ICE-SL (Section 5). Finally, we will offer some concluding remarks (Section 6).

2 English in Sri Lanka: Three circles in miniature

When Sri Lanka became independent in 1948 (under the then official name *Ceylon*), the English language continued to play a significant role as in many other postcolonial contexts in South Asia and beyond. However, the official language policy of the post-independence period was not intended to stabilise English as a medium of communication and instruction – quite on the contrary. The rigid Sinhala-only policy of the 1950s propagated Sinhala as the only official language of the island and denied Tamil, the indigenous language spoken by a substantial minority of the island's population (mainly comprising Tamil-speaking Hindus and Muslims in the Northern and Eastern provinces), an equivalent status. Along these lines, English was not considered to be part of the local linguistic repertoire, and it was not until the mid-1980s, after many years of civil war between Sinhala-speaking and Tamil-speaking Sri Lankans, that English was reintroduced in the Constitution of Sri Lanka as a 'link language' alongside Sinhala and Tamil, i.e. as a neutral interethnic means of communication. In spite of the changing status – and societal role – of English in Sri Lanka over the past six decades, the language remained present on the island all the time and became more and more indigenised: without any doubt, it is true that "Sri Lankan English is not simply 'English in Sri Lanka', but a variety with a certain regional and social identity" (Meshtrie and Bhatt 2008: 200).

However, the picture of English in Sri Lanka or Sri Lankan English as a postcolonial institutionalised second-language variety would be an oversimplification. It is certainly true that for many competent and regular users of English in Sri Lanka, the language is an additional language besides Sinhala or Tamil as their mother tongue. And it is also true that in many structural and functional regards, the variant of English used by those speakers of English as an additional language is a classic case of an emerging New English variety in the Kachruvian 'outer circle'. Unsurprisingly, therefore, it has also been argued elsewhere that in the framework of Schneider's (2003, 2007) dynamic model of the evolution of New Englishes, Sri Lankan English is best characterised by features of nativisation and endonormative stabilisation (cf. Mukherjee 2008). However, as in all other South Asian countries that once formed part of the British Empire, there is also a substantial group of speakers who only display a low proficiency in English and whose usage cannot be regarded as representing an institutionalised variety of English. What is more, there is also a distinct group of native speakers of English in Sri Lanka, including (but by no means being restricted to) the Burgher community (i.e. descendants of European colonists) "for which English is arguably a first language" (Rajapakse 2008: 49). In

essence, then, all three Kachruvian circles – relating to English as a native (ENL), second (ESL) and foreign language (EFL) respectively – are present in Sri Lanka today.

This brings us to a central issue in compiling a representative corpus of English in Sri Lanka: what should ICE-SL capture? Only the language use of 'educated' speakers or exponents of 'standard' English, as clearly defined by Greenbaum (1996b: 6)? But where to draw the line between 'educated' and 'non-educated' English in a New English variety that is still in the process of nativisation and standardisation? While criteria pertaining to the formal English-medium education and to the completion of secondary education may allow for a distinction between competent ESL speakers and low-proficiency EFL speakers, the co-existence of first-language users of English and second-language users, coupled with a much higher normative potential of the former group in spite of the quantitative dominance of the latter group, raises the question of whether or not to include Sri Lankan native speakers of English (ENL) in ICE-SL or not. Our approach in the compilation of ICE-SL so far has been to aim at a reflection of the English language as it is used by competent speakers who meet the ICE standards of age, education and English-medium instruction, regardless of whether they are first or second (or third) language users of English. Thus, we have not explicitly excluded native speakers of English in general or representatives of the Burgher community in particular when selecting the texts for the corpus. Rather than speaking of ICE-SL as an ESL corpus, it would thus be more appropriate to refer to it as a corpus of acrolectal language use by competent English speakers.

The question of what ICE-SL is intended to reflect is certainly one of the most general issues that is also relevant to many (if not all) other ICE components representing New Englishes, but there is a range of further questions referring to ICE-SL-specific methodological challenges and problems that had to be answered in the corpus compilation and annotation process so far. To some of the problems – and the principles of corpus compilation and annotation that we have been following in the ICE-SL project – we will turn in the following section.

3 Principles and problems of corpus compilation and annotation

3.1 Issues of corpus compilation

The ICE project is intended to provide comparable corpora of major first-language and second-language varieties of English world-wide. The comparability of the corpus design across all ICE components is crucial for the validity of all cross-varietal analyses and intervarietal comparisons.

One essential problem in all ICE projects is the emerging time gap between the various ICE components. While ICE-GB, a significant *tertium comparationis* for many ICE-based comparative studies, includes texts from the early 1990s, most texts in the written component of ICE-SL, for example all the student essays, were produced after the year 2000. However, this time gap has to be accepted because it is impossible today to retrieve corpus data from the early 1990s for all the genre categories of ICE-SL. Given that, firstly, the compilation of the written component of ICE-SL covers a period of roughly ten years and, secondly, the collection of spoken data will require several years from 2010 onwards (see Section 5), there is also a diachronic dimension within ICE-SL itself (as in virtually all the other ICE corpora as well). Again, this is a distortion of the synchronic picture of Sri Lankan English that can hardly be avoided.

Even if we accept the inevitablity of certain time gaps between and within various ICE corpora, there are still some implications that cause specific problems in particular genres, especially whenever genres have undergone changes over the past decades. In ICE, this affects in particular the category of social letters: when the ICE corpus design was defined, email communication was virtually non-existent. Over the past fifteen years, this has changed dramatically. Many speakers do not write any social letters in the traditional sense anymore – in many contexts, an informal email is written instead. Obviously, this is a ubiquitous text-linguistic change in English caused by the new written culture of Internet-based communication (cf. e.g. Baron 2004). As in other ICE corpora that are presently being compiled (e.g. ICE-Ghana, ICE-Fiji), we have decided to take account of this genre change and to allow personal emails into the category of social letters. This is not without problems as emails tend to be much shorter on average than traditional social letters. Thus, we have set the minimum length of emails in this category to 50 words, which means that many texts in the text category 'Correspondence' (W1B) are made up of numerous subtexts.

The central question of who counts as a speaker of Sri Lankan English has to be revisited in the context of corpus compilation. As has already been suggested above, the only explicit guidelines concerning the profile of the population to be represented in an ICE corpus refer to "adults of 18+ who have received formal education through the medium of English to the completion of high (secondary) school, but second-language countries might require a university degree" (Greenbaum 1991: 3). ICE-New Zealand, in addition to the above criteria, accepted only the contributions of people who have lived in New Zealand since before the age of 10 and also made restrictions as regards overseas stays (cf. Vine 1999: 10). With regard to extensive stays abroad, however, we refrained

from transferring the restrictive standards of ICE-New Zealand to ICE-SL. Many CVs of Sri Lankan English speakers are characterised by comparatively long stays in ENL (English as a native language) countries such as Australia, New Zealand or the USA. Thus, if one attempts to compile a corpus that is supposed to represent Sri Lankan English, it is essential to include these speakers as well since they form an integral and significantly large part of the English-speaking community in Sri Lanka. Especially fiction writers and academics fall into this category. Note in this context that those speakers who have not spent a considerable amount of time abroad are not free of influence from ENL countries either, because they are frequently exposed to native varieties of English in the English media or in business contexts. Consequently, there generally seems to exist a more or less pronounced ENL influence in Sri Lanka which is not necessarily bound to long-term residence in ENL countries, and it would be a misrepresentation if the corpus data did not mirror this. For that reason, we decided not to adopt the rigid guidelines from ICE-New Zealand in this respect. With regard to author information in this context, it should not go unmentioned that it is extremely challenging to obtain reliable information on the places of residence for most contributors, which further complicates the matter.

3.2 Issues of corpus annotation

After the texts have been collected, structural markup is applied according to the standards as laid down in the markup manual (cf. Nelson 2002). Typographic markup only brings little added value (also in relation to the time and energy that needs to be invested), so we have restricted ourselves to the annotation of unusable characters as regards typographic markup. Concerning content markup, it is the guidelines for quotations that seem to be particularly worth discussing since it is stated that "[i]f quotations comprise one or more complete sentences, mark them *additionally* as extra-corpus text" (Nelson 2002: 11). The reasoning behind this suggestion, which is motivated by the apprehension of the potential inclusion of lengthy passages from speakers that do not use the local variety of English, is unquestionably appropriate for the informational sections of the corpus. However, in the more creative sections, which comprise novels, short stories and press editorials, it is doubtful whether the concept of *quotation* in connection with extra-corpus text can be applied in the same way. In this context, the ICE-SL team (and the ICE-Fiji team which faced similar problems) have decided not to mark quotations which are longer than one sentence as extra-corpus material in the creative sections of our corpora, since these quotations have either been created anew (novels and short stories) or are likely to have been edited (press editorials and press news reports) by a local speaker of English.

The normalisation of the corpus texts is another particularly delicate issue due to the need to strike an adequate balance between making words retrievable on the one hand and keeping normalisation to a minimum on the other. We chose to normalise only those misspellings that are clearly of a typographic nature without a potential grammatical origin (e.g. we would not correct the apostrophe in *the boy's take a walk* because this would suggest that the deviant spelling may not have been caused by the speaker's grammatical considerations).

Another central area of the content markup, namely indigenous and foreign words, is of interest because the difficulties that can arise in the course of the annotation process are presumably not restricted to ICE-SL. Sinhala, an Indo-Aryan language, and Tamil, a Dravidian language with roots in southern India, are the two major indigenous languages of Sri Lanka (cf. Dharmadasa 2007: 116). This means that for ICE-SL, Sinhalese or Tamil words are considered to be *indigenous* while, for example, words of Hindi origin would be categorised as *foreign*. As the linguistic repertoire of the members of the Giessen-based ICE-SL annotation team features neither Sinhala nor Tamil, the temporary tag <in-fo> was introduced as a placeholder tag. This tag marks words which cannot be found in the *Oxford English Dictionary.* Consequently, a word with this tag can either be an indigenous word from Sinhala or Tamil, a foreign word or a Sri Lankan English word. One such difficult case is the word *cädjan* as given in example (1).

(1) The lies flow easily, one after the other, like rainwater pouring off a cädjan roof. <ICE-SL:W2F-020#43:1>

Given that only speakers proficient in Sinhala, Tamil and Sri Lankan English can adequately classify the words concerned, we regularly collect lists with words that have been assigned the tag <in-fo>. These lists are then categorised by Sri Lankans who meet the language competence requirements; in the light of their feedback, the annotation team either replaces the temporary tag <in-fo> with the tags <indig> or <foreign> or simply removes the tag if the word in question is a Sri Lankan English one. In our example, *cädjan* has transpired to be a Sri Lankan English word, which means that it is not considered to belong to one of the indigenous languages, though it might have originated from one of them. It is important to note that competent speakers of Sri Lankan English perceive and use *cädjan* as a lexicalised item of Sri Lankan English. *Cädjan*, with a slightly deviating spelling though, is also listed as a Sri Lankan English word in Meyler's (2007: 46) *Dictionary of Sri Lankan English*. In contrast to Sri Lankan English words, foreign and indigenous words are felt to be closely connected to

their respective languages of origin, which finds expression in the fact that speakers immediately recognise them as foreign or indigenous words in discourse. Examples of foreign words include *bhoota* <ICE-SL:W2D-017#60:1> and *dhãtu* <ICE-SL:W2D-017#9:1> from Pali, an Indian language, while *appochchi* <ICE-SL:W2F-003#73:1> and *goviya* <ICE-SL:W2B-016#20:1> are Sinhalese and thus indigenous words. Our pragmatic approach to the distinction between words belonging to the local variety of English, indigenous words and foreign words, has so far proven to be suitable and efficient.

4 The written part of ICE-SL: Current status and descriptive studies

The first phase of the ICE-SL project, for which basic funding was provided by the University of Giessen and additional funding by the German Academic Exchange Service (DAAD), was targeted at collecting and annotating the 400,000-word written component of the corpus. At present, this first phase is in the process of being completed. The current status of the written component of ICE-SL (February 2010) is summarised in Table 1. As can be seen, all the texts have already been collected, and more than 80 per cent (165 of 200 texts) of the written part of the corpus has already been finalised.

Table 1: Status of the written component of ICE-SL (February 2010)

Category	collected	finalised
Student essays (W1A)	20/20	11/20
Correspondence (W1B)	30/30	20/30
Academic writing (W2A)	40/40	36/40
Non-academic writing (W2B)	40/40	28/40
Press reportage (W2C)	20/20	20/20
Instructional writing (W2D)	20/20	20/20
Persuasive writing (W2E)	10/10	10/10
Creative writing (W2F)	20/20	20/20
TOTAL	**200/200**	**165/200**

While permission to use the texts for ICE-SL has already been obtained for a range of written texts (e.g. student essays and social letters), we will need to initiate a final round of letters asking for – or reminding authors of the issue of – copyright permission upon completion of the written component.

Once the written component has been finalised, it can be used as a corpus of written Sri Lankan English and compared with the written components of other ICE corpora that are already available. It will be particularly interesting to compare the written components of ICE-SL and ICE-GB, the present-day version of the historical input variety, and ICE-India, the neighbouring – and much larger – South Asian variety in order to identify differences between British English and/or the two South Asian Englishes as well as a potential influence from Indian English as an emerging 'epicentre' of English (cf. Leitner 1992) on Sri Lankan English. We also envisage integrating ICE-SL into the set of ICE corpora which we have been using in various studies for the quantitative description of lexicogrammatical differences between various Asian Englishes that represent different stages in the evolutionary process of variety-formation (cf. Mukherjee and Gries 2009; Gries and Mukherjee fc.). It will thus be possible to describe potential correlations between the evolutionary stage of Sri Lankan English and its lexicogrammar against the background of other Asian Englishes.

We have already used pilot versions of the written component of ICE-SL, containing 115 to 130 finalised written texts, for a series of descriptive studies, e.g. on ditransitive verb complementation (Mukherjee 2008), lexical bundles (Deixler 2008) and polysemous verbs such as *give*, *take* and *offer* (Bernaisch 2009; Mukherjee and Werner 2009). Given that some (parts of) genre categories were not included in the pilot versions of the corpus, it will be necessary to see in due course whether the findings from these studies can be replicated for the entirety of the written component of ICE-SL.

5 The spoken part of ICE-SL: Prospects and challenges

The second phase of the ICE-SL project concerns the compilation of the 600,000-word spoken part. The data collection for the spoken component of ICE-SL begins in 2010. Provided that external funding will be available for the second phase of the project, we envisage the completion of the spoken component by 2014. As the sister project ICE-Ghana is also harboured by the University of Giessen (under the auspices of Magnus Huber and his team) and is in almost perfect synchrony with the compilation of ICE-SL, we intend to continue to jointly apply for project funding and to jointly train team members and research assistants in Sri Lanka and Ghana who will be involved in data collection and corpus annotation. The first joint workshop with participants from Sri Lanka and Ghana, funded by the German Research Foundation (DFG), is to take place at the University of Giessen in April 2010.

With regard to collecting the spoken data, we anticipate various challenges. One of these challenges has to do with the various text-types included in the ICE-corpora. Owing to the aforementioned large time-lag between the creation of the first ICE corpora and the newer corpora, some assumptions on corpus design that were accurate in the early 1990s may no longer hold today. Nelson (1996) describes how the selection of text types was intended to fit regions other than Great Britain:

> A corpus dealing exclusively with British English, for example, might include more text types than are represented in ICE. We might wish to include electronic mail messages, faxes and answer phone messages, in order to give a complete view of British English in use in the 1990s. However these text types are not available in all the ICE countries and indeed still have restricted use even in Britain. For these reasons they have been excluded from the general design. (Nelson 1996: 29)

This quotation already hints at some of the problems that are based on the time-lag between the ICE corpora and also sheds some light on the limited extent to which the entirety of English speech communities around the world is captured by the ICE design. A global phenomenon, of course, is the fact that electronic mail has almost entirely replaced social letter writing and to a certain extent also business communication over the last 10 to 15 years.

The same holds true for some of the spoken text-types. In the case of distance conversations, for example, the use of online communication is steadily growing. This new text-type, however, differs from telephone conversations in various respects. Firstly, computer-based distance communication is, at the moment, often planned communication; i.e. the partners in the communicative process prepare for the communication in setting up the communication channel. In telephone conversations, on the other hand, the receiver of the call has often not prepared himself or herself for the conversation. A further difference between classic distance conversation and recent online communication is the possibility to include video. If a video channel is included, online communication resembles direct conversation more than pure audio communication. While such differences in text-type need to be addressed, these new communication channels also offer a range of possibilities for corpus compilation that should not be neglected. Recording classic distance conversations, i.e. telephone calls, requires certain additions to the hardware as recording equipment needs to be set up. Computer-based distance conversation, on the other hand, is a built-in feature of most modern computer systems and conversion from analogue audio formats to digitized formats is no longer necessary. This has some important

advantages for corpus compilation since a larger number of speakers can be included if the ICE team does not have to provide each prospective participant with telephone recording equipment. Furthermore, digital data can be stored more easily; at later stages, it would thus be possible to distribute the source audio data to interested scholars. The issue of data storage and accessibility of source audio for the spoken components of the ICE corpus has so far been largely neglected and it is often not possible for researchers to gain access to the original audio data in question. Note that the transcription of audio data can never feature the entire information of the spoken material.

We expect a further challenge when it comes to collecting some of the public spoken data. While in Great Britain and many other ENL-communities it is relatively easy to gain access to official spoken data such as parliamentary debates or legal proceedings, collecting this data may prove a much larger challenge in Sri Lanka. Although there are public court sessions and public parliamentary debates in Sri Lanka, which is not the case for all ESL-communities, recording and publishing such proceedings may involve certain bureaucratic or even legal problems. In the case of parliamentary debates, bringing recording equipment into parliament or even making notes is not allowed. Although Hansard transcripts are easily available, it has been shown in the past that these transcripts are heavily edited and that they are not a reliable representation of the original spoken performance (cf. Mollin 2007). At present, it remains open whether it is possible to obtain copies of the official Hansard recordings or if it is possible to use television broadcasts of parliament sessions. Matters are further complicated in the case of public spoken discourse owing to the multilingual setting in Sri Lanka. Parliamentary debates, for example, are not conducted in English only but can include Sinhalese, Tamil and English contributions in compliance with Standing Order 12 of the Parliament of Sri Lanka.

The status of English as a link language in Sri Lanka will also play a role in the collection of legal texts as English is not the official language of the Sri Lankan courts. According to Chapter IV, §24(1) of the Sri Lankan Constitution, official languages of the court are Sinhala and Tamil, where "Sinhala shall be used in all the areas of Sri Lanka except those where Tamil is the language of administration".

Apart from the fact that the multilingual and multiethnic situation in Sri Lanka may make it difficult to obtain English texts in some of the categories of the spoken part, this situation has further implications for the choice of speakers that are represented in the corpus. Although the civil war officially ended in May 2009, it is still very difficult to access many areas throughout the island so that the data that can be collected for the corpus may be heavily skewed towards

the Sinhalese part of the population in larger urban areas in the western, southern and central provinces. The numerical dominance of Sinhala native speakers in the Colombo area has recently been shown by Künstler *et al*. (2009), a study on language functions and speaker attitudes in Sri Lanka. Their data show that 84 per cent of their informants were Sinhala native speakers, while only for seven per cent of the speakers the L1 was Tamil (cf. Künstler *et al*. 2009: 59).

A final challenge to be noted refers to the time that is needed to collect and transcribe the spoken data. As has been mentioned above, there is already a diachronic gap between earlier ICE corpora such as ICE-GB or ICE-India that has a negative effect on their comparability. This gap between the sub-corpora is exacerbated by the diachronic gap between the spoken and written components within a single ICE corpus. Nevertheless, partly owing to the challenges discussed in this section, a certain amount of time will be needed to collect the texts for the spoken component, let alone the massive amount of time that the transcription process requires. This largely unavoidable time-lag will affect the synchrony of the spoken and written parts of ICE-SL and the comparability of various ICE corpora. For example, while diachronic gaps of a decade or two are not particularly significant for the description of syntactic structures (as grammatical change is relatively slow), there may be changes in lexis that are captured by newer corpora and corpus parts but not by older ones.

6 Concluding remarks

Despite the large number of challenges and problems involved in corpus compilation and annotation, adding a Sri Lankan component to ICE is of paramount importance with regard to empirical research into New Englishes. The availability of authentic Sri Lankan English material from a wide range of genres allows linguistic research on various structural levels and enables the researcher to critically review conclusions concerning Sri Lankan English which have been drawn on the basis of relatively small sets of data (cf. Herat 2001, 2005). In addition, it has to be emphasised that ICE-SL is the first large-scale representative corpus of Sri Lankan English. As a consequence, ICE-SL can be considered to be the next step in the codification process of Sri Lankan English, which will no doubt improve future lexicographical descriptions of Sri Lankan English. In a wider setting, the compilation of ICE-SL will certainly foster the perception of Sri Lankan English as a variety of English in its own right.

The compilation of ICE-SL is embedded in a network of ICE corpora of the second and third generation that are currently being compiled by ICE teams based in Germany (e.g. ICE-Nigeria in Augsburg, ICE-Malta in Bamberg) and

in Switzerland (ICE-Fiji in Zurich). Owing to the need to exchange information on methodological issues and the goal to increase the comparability of the ICE corpora which are currently being compiled, the ICE teams in Germany and Switzerland have met regularly from 2007 onwards. In various workshops, shared problems regarding corpus compilation and annotation have been discussed and, whenever possible, common standards and guidelines have been set up. We hope that this network will grow and that the emerging standards of corpus compilation and annotation will lead to a high degree of comparabilty across the ICE corpora of the second and third generation.

ICE corpora are limited in size. One million words are often not enough, especially when it comes to the description of low-frequency phenomena in general and the quantitative analysis of linguistic features across genres and corpora in particular. In the light of these limitations, there have been several recent approaches to complement ICE data with other corpora or databases. With regard to corpora of Indian English, for example, Sedlatschek (2009) points out that

> [w]hile closed corpora like […] the Kolhapur Corpus or ICE-India have all the advantages of carefully designed and well researched databases that allow for systematic investigation across registers, modes and text types […], larger corpora are needed to study the use of rare linguistic features. (Sedlatschek 2009: 44)

Following Thelwall's (2005) suggestions, Sedlatschek (2009) opts for the use of online data as a complementary linguistic source by employing the Google Advanced Search Option. In his study of the lexicogrammar of Indian English, the Internet domains that he takes into consideration consist of several Indian and South Asian newspapers that provide free online access. This integrated approach based on carefully designed standard corpora and larger web-derived databases can also be used to investigate other South Asian varieties of English. A comparable research environment for the study of Sri Lankan English is now available at the University of Giessen: in the context of the research project 'Verb-complementational profiles in South Asian Englishes', funded by the German Research Foundation (DFG), very large offline newspaper corpora of South Asian Englishes, including Sri Lankan English, have been compiled which can be used as an additional database for the (quantitative) analysis of acrolectal Sri Lankan English alongside the written part of ICE-SL. It has already been shown in various pilot studies (cf. e.g. Bernaisch 2009) that, with regard to low-frequency phenomena, the integrated analysis of ICE-SL and much larger newspaper corpora of Sri Lankan English is a promising way forward.

Note

1. We are grateful to Dr. Dushyanthi Mendis (University of Colombo) for coordinating the project activities in Sri Lanka and for her invaluable support in the compilation of the corpus. We are also grateful for additional funding provided by the German Academic Exchange Service (DAAD) in 2008–2012 and the German Research Foundation (DFG) in 2010 to support the research collaboration between Giessen and Colombo.

References

Baron, Naomi S. 2004. Rethinking written culture. *Language Sciences* 26: 57–96.

Bernaisch, Tobias. 2009. Verb-complementational nativisation in Sri Lankan English: A pilot study based on standard and web-derived corpora. Giessen: Unpublished Diploma Thesis.

Deixler, Carolin. 2008. Exploring the lexicogrammar of Sri Lankan English: A corpus-linguistic pilot study. Giessen: Unpublished Diploma Thesis.

Dharmadasa, K.N.O. 2007. Sri Lanka. In A. Simpson (ed.). *Language and national identity in Asia*, 116–138. New York: Oxford University Press.

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

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

Greenbaum, Sidney. 1996b. Introducing ICE. In S. Greenbaum (ed.). *Comparing English worldwide: The International Corpus of English*, 3–12. Oxford: Clarendon.

Gries, Stefan Th. and Joybrato Mukherjee. Forthcoming. Lexical gravity across varieties of English: An ICE-based study of speech and writing in Asian Englishes. *International Journal of Corpus Linguistics*.

Herat, Manel. 2001. Speaking and writing in Lankan English: A study of native and non-native users of English. *California Linguistic Notes* 26(1). Available at http://hss.fullerton.edu/ linguistics/CLN/spring01_articles/herat.pdf. Accessed on 16 Jun 2009.

Herat, Manel. 2005. BE variation in Sri Lankan English. *Language Variation and Change* 17: 181–208.

Künstler, Victoria, Dushyanthi Mendis and Joybrato Mukherjee. 2009. English in Sri Lanka: Language functions and speaker attitudes. *Anglistik – International Journal of English Studies* 20 (2): 57–74.

Leitner, Gerhard. 1992. English as a pluricentric language. In M. Clyne (ed.). *Pluricentric languages: Differing norms in different nations*, 179–237. Berlin: Mouton de Gruyter.

Meshtrie, Rajend and Rakesh M. Bhatt. 2008. *World Englishes: The study of new linguistic varieties*. Cambridge: Cambridge University Press.

Meyler, Michael. 2007. *A dictionary of Sri Lankan English*. Colombo: Meyler.

Mollin, Sandra. 2007. The Hansard hazard: Gauging the accuracy of British parliament transcripts. *Corpora* 2 (2): 187–210.

Mukherjee, Joybrato. 2008. Sri Lankan English: Evolutionary status and epicentral influence from Indian English. In K. Stierstorfer (ed.). *Anglistentag 2007 Münster: Proceedings*, 359–368. Trier: WVT.

Mukherjee, Joybrato and Stefan Th. Gries. 2009. Collostructional nativisation in New Englishes: Verb-construction associations in the International Corpus of English. *English World-Wide* 30 (1): 27–51.

Mukherjee, Joybrato and Janina Werner. 2009. Highly polysemous verbs in New Englishes: A corpus-based pilot study of Sri Lankan and Indian English. Paper presented at the 30th ICAME Conference, Lancaster, 27–31 May 2009.

Nelson, Gerald. 1996. The design of the corpus. In S. Greenbaum (ed.). *Comparing English worldwide*, 27–35.

Nelson, Gerald. 2002. *Markup manual for written texts*. London: Survey of English Usage.

Rajapakse, Agra (2008). A descriptive analysis of the language of the Burghers of Sri Lanka. In D. Fernando and D. Mendis (eds.). *English for equality, employment and empowerment*, 49–58. Colombo: SLELTA.

Schneider, Edgar W. 2003. The dynamics of new Englishes: from identity construction to dialect birth. *Language* 79(2): 233–281.

Schneider, Edgar W. 2007. *Postcolonial English: Varieties around the world*. Cambridge: Cambridge University Press.

Sedlatschek, Andreas. 2009. *Contemporary Indian English: Variation and change*. Amsterdam: John Benjamins.

Thelwall, Mike. 2005. Creating and using web corpora. *International Journal of Corpus Linguistics* 10 (4): 517–541.

Vine, Bernadette. 1999. *Guide to the New Zealand component of the International Corpus of English (ICE-NZ)*. Wellington: Victoria University.