Is it Creole, is it English, is it valid? Developing and using a corpus of unstandardised written language.

Mark Sebba and Susan Dray
Department of Linguistics and Modern English Language, Lancaster University,

This paper will present our experiences of developing and using two computer corpora of written Creole. By ‘Creole’ here we mean English-lexicon creoles of Caribbean origin, which are also used in Britain. Creole, both written and spoken, is unstandardised and subject to a high degree of variability at all levels – especially in grammar, phonology and orthography - because of its relationship with Standard English.

The corpora we will discuss are part of an on-going project at Lancaster University, investigating the practices of writers using Creole. They are:

- the Corpus of Written British Creole (CWBC), a collection of texts written wholly or partly in Creole by West Indians whose formative years have been spent in Britain

- a Corpus of Written Jamaican Creole (CWJC), a collection of texts of diverse types written by Jamaicans in Jamaica

We will briefly discuss issues which arose when setting up the corpora, in particular the practical issue of the identification of texts for inclusion, and the specific problems that this raises with respect to English-lexicon Creoles. Can a text containing Creole features be included in the corpus if the writer’s target appears to be Standard English? Where Creole occurs together with Standard English, to what extent is it necessary to include the Standard English (SE) parts in the corpus – and where is it legitimate to make a ‘cut’?

We then focus on some potentially complex linguistic questions which arose during the annotation procedure, e.g. to what extent do graphological features, such as punctuation, layout, and the use of upper and lower case letters in a text, form part of a ‘naturally’ developing orthography? Is it important for these features to be indicated in the corpus? To what extent do an individual’s literacy skills affect what is written, and what are the implications of this for the representativeness of the corpora?

We will end with some remarks on the potential of these corpora as a research tool in order to illustrate the application of the annotation method. Linguistic items and their orthographic representations can be compared and contrasted with variables such as demographic information (about writers or voices within texts), country, text type, and social context. This could provide information, for example, not only on orthographic variations within or across texts, but also on how writing practices may vary according to social, cultural, generational and educational factors as well as over time.