In their own voices, in their own words: A sociolinguistically-informed corpus of Nigerian Arabic

The sample

The texts directly reflect the sociolinguistic basis of the research which informed their collection (Owens 1998). This means that the texts divide into a number of subgroups, which include the following. The numbers indicate how many texts were recorded for each aspect of the study, and doesn't reflect the number which are currently available.

Maiduguri interviews (N = 56) (abbreviated, IM). The main focus of the sociolinguistic study was a characterization of Arabic in Maiduguri (500,000 – 1,000, 000 inhabitants) from a variationist sociolinguistic perspective. To this end, the first set of recordings were interviews carried out over four basic demographic groups in Maiduguri: men and women x over 50, under 35. A third parameter which in terms of cell count is only partially balanced, is area of residence in Maiduguri: Ruwan Zafi, Gambori, Gwange or elsewhere.

1. Village interviews (originally, N = 52) (abbreviated TV). As matters turned out, it quickly became clear that understanding the current situation in Maiduguri required a good knowledge of the 'ancestral' dialectal features of the new urban inhabitants. To this end the interview format was extended to villages in all regions of Arabic-speaking Nigeria, excluding diaspora populations in other Nigerian cities (e.g. Lagos).

2. Free group conversations (abbreviated GR). To get an impression of non-interview speech, the Maiduguri data was further enhanced via a number of open-ended conversations. Speakers were set together and their conversation recorded.

3. Codeswitching texts (N = 10) (abbreviated CS). As research progressed it became an interesting question to consider Maiduguri Arabs not only as speakers of Nigerian Arabic, but in terms of their total linguistic repertoires. All Maiduguri Arabs under the age of 50 are bilingual at least in Hausa, sometimes multilingual in English and Kanuri as well, and many have some background in Standard Arabic. In the mid 1990's the research was extended to consider this aspect of language as well.

4. Other texts. There are further texts of various sorts, e.g. a recording of an Arabic radio show, which will be posted.

The texts themselves are presented in two different formats. First there are texts totalling about 10 hours in total which have audio and transcription, and which have been translated into English and annotated for major linguistic and cultural characteristics. Secondly there are texts with audio and transcription, but no translation. Altogether some 30 hours of audio and accompanying transcriptions now available. While the public corpus is not yet finished, further progress, both in respect to the addition of more texts and to translation and annotation of existing ones, will depend on the availability of time and resources.

Technical problems: Not all audio and digital data has survived. Although the transcriptions were all digitalized in the early 1990's, in the course of transferring the material from DOS to later Windows formats, some of the transcriptions were either lost altogether or were compromised in some way.

A note on translation. The translation is far closer to the idiomatic end of the scale than the literal. There has been little attempt to keep to the Arabic grammatical structures in the English translation and false starts and other disfluencies have generally been disregarded. Occasionally literal meanings are indicated in parentheses in the translations, or in footnotes, though no attempt has been made to be consistent in this respect. It is intended in the future to provide a part of the texts with interlinear, morphologically segmented glosses, where the literal structure will be evident.