2006-06-01

DIT Speech Corpus

Dermot Campbell
Dublin Institute of Technology, dermotfcampbell@gmail.com

Yi Wang
Dublin Institute of Technology, yi.wang@dit.ie

John Kelleher
Dublin Institute of Technology, john.d.kelleher@dit.ie

Marty Meinardi
Dublin Institute of Technology, marty.meinardi@dit.ie

Bunny Richardson
Dublin Institute of Technology, bunny.richardson@dit.ie

DIT Speech Corpus
2006
IVACS
Nottingham

DIT's nascent speech corpus will allow a body of spoken material to be searched for features of informal native speech via a normalised transcription. Once located, the original sound files can be played at normal speed or slowed down in order to better study the recorded speech.

The DIT speech corpus treats speed of delivery as a key element in producing the elisions, assimilation, reductions and co-articulations characteristic of native-to-native dialogues. Lack of training in dealing with this spoken register can lead to lack of preparation for the world of real speech and even to a degree of social exclusion.

It is also envisaged that non-native speech will be included in the corpus so that comparisons can be drawn between native speech and that of various nativised productions of the same items. The database will therefore be capable of being queried on a multi-factorial basis depending on user needs.

The optimal segmentation of the normalised transcript is, however, far from clear, and some of the difficulties will be touched on by this presentation. While the tone unit, as proposed by David Brazil, for example, is attractive as a base unit for displaying the concordanced speech corpus, it nevertheless raises problems when there is a discrepancy between semantic segmentation and actual phonetic delivery.

The rationale for the currently adopted minimal unit will be explained and members of the audience will be invited to offer feedback on any requirements their own use of corpora would place on the database.

Keywords: speech corpus, speaking speed, transcription, tone units