2009-9

DIT's Dynamic Speech Corpus

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

Ciaran McDonnell
Dublin Institute of Technology, ciaran.mcdonnell@dit.ie

Marty Meinardi
marty.meinardi@dit.ie

Charles Pritchard
Dublin Institute of Technology, pritchard@dmc.dit.ie

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

See next page for additional authors

Authors
Dermot Campbell, Ciaran McDonnell, Marty Meinardi, Charles Pritchard, Bunny Richardson, and Yi Wang

Digital Media Centre

Conference papers

Dublin Institute of Technology *Year* 2009

DIT's Dynamic Speech Corpus

Dermot Campbell* Ciaran McDonnell† Marty Meinardi‡
Charles Pritchard** Bunny Richardson†† Yi Wang‡‡

*Dublin Institute of Technology, dermotfcampbell@gmail.com
†Dublin Institute of Technology, ciaran.mcdonnell@dit.ie
‡Dublin Institute of Technology, marty.meinardi@dit.ie
**Dublin Institute of Technology, pritchard@dmc.dit.ie
††Dublin Institute of Technology, bunny.richardson@dit.ie
‡‡Dublin Institute of Technology, yi.wang@dit.ie

— Use Licence —

Attribution-NonCommercial-ShareAlike 1.0

You are free:

- to copy, distribute, display, and perform the work
- to make derivative works

Under the following conditions:

- Attribution.
  You must give the original author credit.
- Non-Commercial.
  You may not use this work for commercial purposes.
- Share Alike.
  If you alter, transform, or build upon this work, you may distribute the resulting work only under a license identical to this one.

For any reuse or distribution, you must make clear to others the license terms of this work. Any of these conditions can be waived if you get permission from the author.

Your fair use and other rights are in no way affected by the above.

DIT's Dynamic Speech Corpus

Campbell, D., McDonnell, C., Meinardi, M., Pritchard, C., Richardson, B., Wang, Y.

The Digital Media Centre of the Dublin Institute of Technology undertakes applied, multi-disciplinary research with the help of external funding. The ***FLUENT*** project outlined below, which is funded by Enterprise Ireland, involves the construction of a Dynamic Speech Corpus (DSC). This is a resource aimed mainly at learners of English, but is sophisticated enough to also address the needs of teachers, authors and researchers.

Speech Corpus or Spoken Corpus?

The DSC is deliberately called a ***speech*** corpus. This is to distinguish it from existing ***spoken*** corpora, which study the form of spoken language and study that which has been spoken. The ***FLUENT*** DSC, on the other hand aims at making the act of speech production itself available to learners and researchers. It is not the transcript of spoken language which is important, but the actual sound files themselves and those findable, reduced features of spoken language which are the subject of study.

Most current spoken corpora use readily available speakers (e.g. students) in accessible situations (e.g. seminar presentations) and are recorded so as to maintain 'naturalness'. But in many recordings, 'naturalness' equates to a low audio quality, e.g. telephone recordings, ambient noise, or a messy signal. Other recordings have been made with speech synthesis in mind and therefore be totally unsuitable for learning purposes.

In contrast, the DSC uses industry-standard recording techniques while retaining a high degree of naturalness. The unscripted dialogues it contains are similar to telephone conversations between friends, but with an audio standard that can bear instrumental analysis.

Traditional Teaching Dialogues

Dialogues written for classroom use are characterised by short, self-contained, focused interchanges which are politely 'choreographed'. Speaker A finishes a turn completely before Speaker B takes up his/her turn. There is rarely any cross talk or back-channelling. The aim of these 'dialogues' is to increase the learner's vocabulary in a coherent (realistic) context and to demonstrate correct application of linguistic structures. They can be good production models for L2 speakers of the language, but they are inadequate for promoting dialogic fluency. They are like a series of interleaved monologues rather than L1-L1 dialogues and do not represent the way L1 speakers actually interact.

Real Dialogues

Genuine dialogues, on the other hand, do not exist in order to demonstrate anything, but rather to realise a communicative goal. We rarely speak for the sake of speaking, but rather to influence our interlocutor, effect a change, achieve a goal, etc. There is a purpose towards which we steer our listener. In fact, for every speaker there are two listeners: the interlocutor and the speaker him/herself. In genuine dialogues speakers monitor and adjust their speech production in light of the development of the dialogue. It is a highly interactive process and fluency in this context consists not in a *legato*, coherent flow of speech characterised by syntactic elegance, but rather a 'confluency' of two speakers.

McCarthy and Tao (2008) have looked at the importance of appropriate turn-taking with regard to fluency. They propose that in order for speakers to be deemed 'fluent', they need to be 'confluent', i.e. they need to be able to interact naturally. In order to do this they highlight three important features of natural dialogue: chunks, linking items and 'small words'. When interlocutors do not use these items, the dialogue sounds unnatural.

Ready-made chunks, such as: *you know*, *I mean*, *what do you think*, and *or something like that*, have an interactive function 'connecting, as it were, the speakers together'. Tao (2003) found that items that link to the previous turn are the norm (e.g. *uh-uh*, *yeah*, *well*, *right*) while items which do not are rare. Without such linking, flow between turns is disrupted. The linking items also allow for thinking time or pause-time just after them, so they may be placed immediately the previous speaker finishes, without silence or over-hesitation between turns.

What Hasselgreen (2004) calls small words (*well*, *actually*, *cos*, *just*, *so*, *like*) have high frequency in any L1 conversational corpus, but a much lower frequency in written corpora. They have an important interactive function.

If we look at an unscripted dialogue, the occurrence of McCarthy and Tao's 3 confluency items becomes clear. What becomes apparent in a natural unscripted L1 to L1 dialogue is the structured messiness. There are few, if any, complete phrases, there is a lot of overlap and cross-talk between interlocutors, yet the flow is not interrupted; in fact it flows better.

Speech and the Written Word

Flowing L1-L1 speech can be compared to a signature, where the individual letters of the name, middle name and surname are often indistinct – the three elements may even be run together and blurred. This is similar to L1 informal speech where speakers use the minimum of effort to articulate. The initiative resides with the speaker. If the listener cannot understand the speaker under these circumstances, then the speaker is obliged to do a 'second pass' and the needs of the listener are highlighted. The speaker is obliged to articulate more carefully in order to achieve intelligibility. Following the writing model, we still have handwriting, but now the words are separated and each letter (phoneme) is distinguishable. Finally, careful, broadcast speaking could be compared to the printed word, where each letter, let alone each word, is in citation form.

What learners find difficult to understand is that there are no words in speech, only a speech continuum. The words are in the heads of the interlocutors, not in the speech itself, and communication is successful when the listener is able to attribute the correct lexical items to the relevant sequences of the speech signal.

The DSC audio recordings are accompanied by idealised, orthographic transcripts. This allows the user to understand the semantic content of the lexical items in the speech flow and contrast the clarity of the written version with Cauldwell's (2002) 'messiness' of real speech. The learning effect is in the comparison of the speech which the transcript triggers in the learner's head (which will be different in each individual case) and the sequences actually spoken by the L1 speakers. The idealised transcript also allows all occurrences of a search string to be retrieved (from hyper-articulated to hyper-eroded), listened to and compared.

Spend more Time with the Signal – Literally!

Cauldwell (ibid) urges us to spend more time studying ***how*** something was said, rather than ***what*** was said, and here again the DSC obliges. Each speaker in a dialogue can be heard in isolation, or faded in/out so that the dialogue can be followed while concentrating on one of the interlocutors. Each segment can be listened to at normal speed or slowed to anything down to 40% of normal speed – without tonal distortion. This means that the natural prosody of real dialogue can be studied, as it were, in slow motion, but without the tonal shifts associated with physically slowing a recording. Just as the high-speed filming of a tennis serve can – when slowed down – focus attention on the snap of the wrist at the point of contact with the ball, so too the slow-down technique allows attention to be paid to the ***manner*** of speech production rather than to the content of what was said.

Mehrabian (1967) estimated that a full 38% of successful, informal communication, where personal attitude of the speaker is involved, is due to the manner in which speech is produced, rather than the choice of word, which accounts for only 7%, with the remaining 55% due to facial expression, gestures, feedback, etc. The application of time-scaling allows attention to be focused on that 38%. At slow speeds, such as the 40% practical limit used in the DSC, blur, elision, assimilation, coarticulation, changes in pitch direction, vowel lengthening, and so forth are highlighted in a way that is not possible at normal speeds of delivery. There is a *Verfremdungseffekt* – a distancing effect – which applies when normal speech is slowed down by a factor of 2 to 2.5. This is similar to looking at an optical illusion. The brain cannot focus on both interpretations built into the picture (e.g. the girl or the flowers) at the same time. In a similar fashion, listening to speech which has been slowed down allows prosody and intonation patterns to be foregrounded and the semantic content of the words played down.

Aims of the Dynamic Speech Corpus

The DSC is a tool which can be used in conjunction with any course materials to prepare students to work or live in an L1 speech community. Since it provides an orthographic, idealised transcript, and since each communicatively significant feature is tagged, it is possible to find samples of speech features being studied by means of multivariant searches. The database can be searched by text string or linguistic feature (e.g. speaker intention, formulaic sequences, turn behavior, expressivity etc.) and the samples found, listed in a concordanced view. These can be clicked on in turn to play-and-contrast the

various examples returned. Each sample can then be listened to in slow-down mode; or the dialogic environment which gave rise to the sample can be accessed and the pragmatics of the speech production studied. How the string was said, by whom, in response to what, and by way of turn taking, turn retention or turn contention are all dynamic features of speech which can be made accessible to the user due to the architecture of the DSC.

Since the linguistic provenance of each speaker is indicated in the DSC, it is possible to look for US English samples, IRL, UK, ZA or any other L1 English variety, and as the DSC becomes populated over time, it can also increase in diachronic value.

From 'Battery fed' to 'Free-range': Ways into the DSC

The recordings in the DSC are unscripted interchanges between L1 speakers of several English varieties. The dialogues contain samples of L1-L1 reductions which can be found via multivariant searches, played and contrasted and then the semantic and phonetic environment in which they were uttered studied at normal or slowed speeds.

While this sort of resource is suitable for advanced learners or researchers, the DSC could also be approached in a scaffolded manner, provided that materials writers adapt current offerings to take advantage of its features. At the low end of the linguistic 'food-chain' are the scripted dialogues, discussed above, which are necessary, but insufficient to train learners to survive in an L1-L1 environment. These are the 'battery fed' dialogues contained in all learning materials.

A step up from that are the 'corn-fed' dialogues of story-boarded but unscripted interchanges. Here speakers are given roleplay guidelines of a course-book scenario, so that there is direction to the dialogue, but the speech still exhibits a degree of individual disfluency while maintaining dialogic fluency.

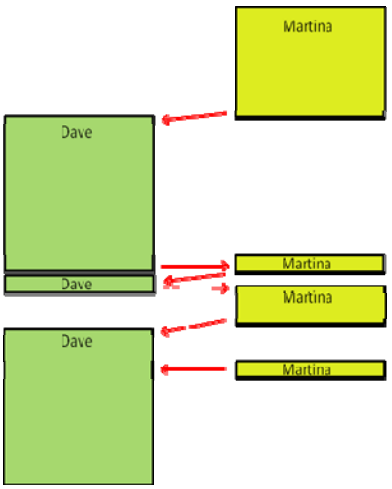
The third and last step is entry into the 'free-range' resources of the DSC itself. The speakers are well known to each other and relax quickly in the course of the recording, so that the reduction features of L1-L1 informal dialogue can be captured at a high audio quality and without 'natural' but signal-degrading extraneous noise.

It is proposed to include a sample of this three-stage approach in training materials accompanying the DSC. Given the availability of the DSC it will require only a minor adjustment in current training materials to prepare students to avail of resources which cannot realistically be included in current teaching/learning materials.

The DSC and Self-study Mode

The 'new learning paradigm' shifts the emphasis from teaching to learning. An introductory section to the DSC will alert the user to the sorts of unexpected speech features available in the corpus and how to find them. This will permit the learner to use the resource in self-study mode and free up precious class time for targeted teacher interventions – a level of interactivity that no self-study resource can (or should be expected to) match.

One of the main insights the DSC can afford the learner is the dynamics of turn-taking in real dialogues. Real speakers are not just 'serial speakers', but often speak at the same time (the DSC can separate each speaker for individual study, while maintaining the naturalness of the interchange), backchannel, add throw-away comments, etc. Turn-taking behaviour is flagged in the DSC and turn construction, turn maintaining, turn contention and turn relinquishing strategies can be studied in a principled fashion.



Key to L1-L1 fluency are small elements such as chunks, linking words and 'small' words. The phonetic environment in which these are uttered means that speakers often deliver their communication in phonetic and expressive 'envelopes' as short as 3 to 4 words long, before pausing, changing pitch or speed of delivery. None of these communicative features can be studied via a transcript and therefore a principled access to them via the audio assets is necessary – and available in the DSC. Users will also be able to search the corpus on a particular topic (e.g. travel), slow down the speech to study its prosody, study the phonetic characteristics of connected speech, find similar samples spoken at different speeds, or find strings spoken with different levels and manners of expressivity. The *FLUENT* project finishes in June 2010.

Dermot Campbell is Head of Dept. of Applied Languages at DIT, currently seconded as a full-time researcher in the Digital Media Centre (DMC). He is mainly interested in the development and application of speech technologies.

Ciaran McDonnell is Head of Research at the Digital Media Centre. He has supervised 10 PhD students.

Charlie Pritchard is Manager of the DMC. He has managed over 30 major projects at national and EU level and strongly encourages multi-disciplinary, applied research.

Marty Meinardi is a lecturer in the School of Languages at DIT and is also a researcher on the FLUENT project. She is particularly interested in the use of authentic English in teaching.

Bunny Richardson is completing her PhD in the DMC and working full-time on the FLUENT project. Her area of study is English for International Communication

Yi Wang is a final year PhD student in the DMC. She is particularly interested in formulaic sequences and the application of DMC technologies to the acquisition of English intonation patterns by Chinese learners.

dermot.campbell@dit.ie

References

Aitchison, J. (1994). Understanding words. In: G. Brown, K. Malmkjær, A. Pollitt and J. Williams (Ed.). *Language and Understanding*, Oxford: Oxford University Press

Brown, G. (1990). Listening to Spoken English. 2nd ed. Harlow: Longman

Brown, G. & Yule, G. (1983). Teaching the Spoken Language: an approach based on the analysis of conversational English. Cambridge: Cambridge University Press.

Campbell, D.F., Wang, Y., McDonnell, C. (2008). FS ≠ FS (Formulaicity and Prosody), BAAL 2008, Swansea, UK

Campbell, D.F., Wang, Y., McDonnell, C. (2007). A Prototype Speech Corpus, EuroCALL 2007, Coleraine, N. Ireland

Carter, R. & McCarthy, M. (1997). Exploring Spoken English. Cambridge: Cambridge University Press.

Cauldwell, R. (2002). Phonology for listening: relishing the messy, www.speechinaction.net

Cruttenden, A. (1997). Intonation, Cambridge: Cambridge University Press

Cullen, R. (2001). 'PPP and Beyond: Towards a Learning-Centred Approach to Teaching Grammar'. In Ferrer Mora, H. et al. (eds.). 2001. Spain: Universidad de Valencia.

Field, J. (1998). Skills and strategies: towards a new methodology for listening. *ELT Journal*, *52 (2):* 110-118

Field, J. (2003). Psycholinguistics. A resource book for students. London: Routledge

Gibbons, P. (2002). Scaffolding Language, Scaffolding, Learning: Working with ESL Children in the Mainstream Elementary Classroom. New Hampshire, USA: Heinemann.

Hasselgreen, A. (2004). Testing the spoken English of young Norwegians: A study of test validity and the role of 'smallwords' in contributing to pupils' fluency. Cambridge: Cambridge University Press.

McCarthy, M. (2006). Fluency and confluence: what fluent speakers do. In: M. McCarthy, (Ed). *Explorations in Corpus Linguistics.* Cambridge: Cambridge University Press. Ch.1.

McCarthy, M. and Tao, H. (2008). Profiling spoken fluency. Feeling like an independent user of English. *The Language Teacher* 32.07 July 2008)

Maybin, J., N. Mercer and Steirer, B. (1992). " 'Scaffolding' Learning in the Classroom". In Norman, K. (Ed.). *Thinking Voices: The Work of the National Curriculum Project.* London: Hodder and Stouton for the National Curriculum Council.

Mehrabian, A. and Wiener, M. (1967): Decoding of inconsistent communications. In *Journal of personality and social psychology* 6(1): 109–114.

O'Keeffe, A., McCarthy, M., Carter, R. (2007). From Corpus to Classroom, Cambridge: Cambridge University Press

Sinclair, J. McH. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Tatham, M. & Morton. K. (2006). Speech Production and Perception. New York: Palgrave Macmillan.

Tao, H. (2003). Turn initiators in spoken English: a corpus based approach to interaction and grammar. In P. Leistyna & C. Meier (Eds.), *Corpus Analysis: Language Structure and Language Use* (pp. 187-207). Amsterdam: Rodopi.

Walsh, S. (2006). Investigating Classroom Discourse. London: Routledge.