2007-01-01

# The Need for a Speech Corpus

Dermot Campbell
*Technological University Dublin*, dermot.campbell@tudublin.ie

Ciaran McDonnell
*Technological University Dublin*, ciaran.mcdonnell@tudublin.ie

Marty Meinardi
*Technological University Dublin*, marty.meinardi@tudublin.ie

Bunny Richardson
*Technological University Dublin*, bunny.richardson@tudublin.ie

## Recommended Citation

Campbell, D. et al. (2007) The Need for a Speech Corpus. *ReCALL,* vol.19, pt. 1, January. doi:10.1017/S0958344007000213

# *The Need for a Speech Corpus*

DERMOT F. CAMPBELL
*Digital Media Centre, Dublin Institute of Technology, Ireland*
*(e-mail: dermot.campbell@dit.ie)*

MARTY MEINARDI
*School of Languages, Dublin Institute of Technology, Ireland*
*(e-mail: marty.meinardi@dit.ie)*

BUNNY RICHARDSON
*School of Languages, Dublin Institute of Technology, Ireland*
*(e-mail: bunny.richardson@dit.ie)*

CIARAN MCDONNELL
*Digital Media Centre, Dublin Institute of Technology, Ireland*
*(e-mail: ciaran.mcdonnell@dit.ie)*

## Abstract

This paper outlines the ongoing construction of a speech corpus for use by applied linguists and advanced EFL/ESL students.

The first section establishes the need for improvements in the teaching of listening skills and pronunciation practice for EFL/ESL students. It argues for the need to use authentic native-to-native speech in the teaching/learning process so as to promote social inclusion and contextualises this within the literature, based mainly on the work of Swan, Brown and McCarthy.

The second part addresses features of native speech flow which cause difficulties for EFL/ESL students (Brown, Cauldwell) and establishes the need for improvements in the teaching of listening skills. Examples are given of reduced forms characteristic of relaxed native speech, and how these can be made accessible for study using the Dublin Institute of Technology's slow-down technology, which gives students more time to study native speech features, without tonal distortion.

The final section introduces a novel Speech Corpus being developed at DIT. It shows the limits of traditional corpora and outlines the general requirements of a Speech Corpus. This tool–which will satisfy the needs of teachers, learners and researchers–will link digitally recorded, natural, native-to-native speech so that each transcript segment will be linked to its associated sound file. Users will be able to locate desired speech strings, play, compare and contrast them—and slow them down for more detailed study.

## 1 The use of authentic language in ELT; can a speech corpus make the difference?

A long-standing argument in academic EFL/ESL circles is that learners of English will most likely converse with other Non-Native Speakers (NNSs) of English, and that they will therefore not need to be able to use the discourse structures and values of a Native-Speaker (NS). However, learners of English will, and do have transactional encounters with NSs and they will have to be able to at least understand the native-speaker's discourse pattern, even if the learner is not linguistically equipped yet to reciprocate in this discourse pattern and join in to build a relationship with the interlocutor. It is also important not to disregard those learners who choose to live and work in the country of their chosen L2 and who are therefore more likely to have dealings with NSs. Confidence in being able to repair a misunderstood message purely from its context is a skill NSs use continuously in communication the ability to rely on contextual and cultural cues however, is something which NNSs have to re-learn in a foreign language. It is therefore important to expose the NN learner to as much authentic language in the L2 as possible in order for the learner to be able to build up a sufficient store of contextual and cultural clues. Whereas both written and spoken authentic materials are to be recommended, the argument in this article shall be made in favour of exposure to natural, spoken material which will most benefit the learner listener and prepare him/her for encounters with the L2 in day-to-day living. Although authentic audio material may previously have been deemed unsuitable for language learners because of the difficulties spoken NS speech poses, the present authors aim to demonstrate that through a Speech Corpus and concordancer teamed with the added benefit of the *DITCall* slow-down tool, 'real' NS language can be made accessible to the language learner. The unique variable slow-down facility, without pitch distortion, for speech recordings allows students to capture details in native, natural spoken English collected in the Speech Corpus.

## 2 Perceived difficulties in using authentic language learning materials

Researchers such as Bacon and Finneman (1990) and Herron and Seay (1991) report on the cognitive and affective benefits for learners who work with oral authentic texts and note that explicit attention to the development of listening skills improves listening comprehension at all levels of instruction with no negative effect on grammar, vocabulary, or oral skills. Bacon and Finneman furthermore observe that authentic oral and written input not only facilitates learners in relating form to meaning but also motivates students as it helps to create shared knowledge and understanding between NNSs and NSs.

However, despite these positive findings the use of authentic spoken language is neither uncontroversial nor widely used in EFL/ESL teaching. The main perceived difficulties with authentic language for second language learning are a possible lack of context and - for oral language - the speed and consequent idiosyncrasies of spoken NS language. In his reply to Carter, Cook (1998: 61) for example warns of the drawback of authentic spoken language: 'A good deal of actual language use is inarticulate, impoverished and inexpressive.' It seems essential to separate the use of authentic input data for listening purposes and those authentic data presented to the learner for emulation within the discussion on the use of authentic materials. It is clear that learners may want to choose their own individual form of spoken English language, whether influenced by the L1 or whether the NNS chooses to produce a good interpretation of a native-like accent. The learners' goal of wanting to be able to understand natural NS English, spoken at speed, and their productive goal of wanting to be able to communicate in the foreign language (at whatever level they themselves choose) should therefore not be seen as one and the same. It is felt that Cook's warning is therefore only relevant to the teaching of productive skills.

Similar criticism has been expressed about the use of corpora (especially spoken corpora) in language teaching. Tribble (1997) notes the possible methodological problems that large quantities of language corpora may pose for language learners and teachers alike. And in the debate about the authenticity of corpus data, researchers such as Widdowson (2000), Cook (1995) and Mishan (2004) feel that once a recording is transcribed for analysis the material could no longer be seen as authentic, as it was taken out of its natural habitat. Although this seems a valid point to make, the counter argument is that of the educational value of being an observer of discourse, rather than a participant (this is after all a regular occurrence in real life). It must also be remembered that for some learners of English the classroom setting and, more importantly, the classroom language could not be further removed from the reality of the native speakers whose language they are trying to learn. It is a well known fact that there are still many English classrooms in Asia for example, where the students never get access to English NS speech and where even the teachers do not use spoken English in the classroom. One could argue that there are possibilities of including realia and other 'authentic' multimedia texts from the target country into the lesson material, but this does not make up for the fact that the students have never been exposed to real native speaker speech in realistic contexts. Crystal (1981:90, 91) is of the opinion that it is essential for learners of English to be exposed to 'real' conversational English and feels that very often, language learning material does not manage to bridge the gap between 'classroom language and language in use'. He believes that real everyday conversation is the type where we are not on our 'best linguistic behaviour' and that learners need to be able to identify with that type of language, in order to become part of the target language community.

## 3 Using corpora in lesson material

Finding the right model of spoken English for the use of language learning material that is appropriate for the target group of learners is certainly not an easy task. How does one use the 'garbled' spoken NS speech material in such a way that it is accessible to the leaner, without having to resort to either editing the contents or scripting the original material, so that it can be re-worked with the use of voice actors? With the development of the Speech Corpus and the use of the *DITCall* slow-down algorithm, the door seems open to using formerly inaccessible material, or material which was too difficult for learners to process because of its speed. The algorithm will allow recorded speech to be slowed down to any desired speed in the range of 0.5 to 1.0 times normal speed without affecting the pitch.

Crystal (1981: 92, 93) notes that 'controlling' texts for the use of language learning textbooks, renders them unrealistic and un-useful to the learner and he calls attention to the fact that 'real' speakers make mistakes, are interrupted, use vague language, hesitate, argue etc. and that the language the student is confronted with within the target community '…is a fundamental change from a pedagogically oriented world, in which people make allowances for mistakes and incomprehension, to a world of quite a different character'. The benefit of using corpus data is that it can reveal interesting grammatical structures and discourse patterns in spoken English and is able to convey information about discourse genre and socio-cultural values of the source of the data. A corpus with spoken audio data can expose language content, which is appropriate to the learner's (either current, or projected) geographical surroundings. The inclusion of other NS varieties of English in a corpus can provide alternative NS models for learners and linguistic scholars and show up interesting linguistic deviations from the British English norm. The practicalities of the characteristics of connected speech, which cannot be shown in written corpus texts, can be highlighted through the use of a spoken corpus, which benefits the learner's receptive skills. In particular a concordance of spoken language can show the learner the difference in pronunciation of a word depending on its occurrence in a sentence. Unlike text corpora, a spoken corpus can reveal prosodic information and also shows preferred pronunciation styles of the chosen socio-cultural model.

It can be argued that exposing the learner to a type of language material which will be as closely modelled on reality as possible will make the learning process more relevant to the learner and will therefore help motivate the student to improve their skills in the target language. If a learner feels that they will be able to use the language learned in the 'real world', the efforts concerned in the learning process will seem more of an investment rather than a chore. One can imagine that especially adult learners will only use their acquired skills if they feel that their efforts will be appreciated and understood by other English speakers; in other words: that they will not look silly using their newly acquired language. If learning materials have seemingly no relevance to the learners' own reality – and prospective reality, for example for those learners who will also attempt to live in the country of the target language – there will be very little motivation on the learners' part to improve on their language skills.

## 4 Integration into the speech community using the speech corpus

It can be envisaged that for those learners of English who are learning the language in the target country itself, spoken corpora recorded in the immediate environment of that student, and presented in both transcribed and audio form, can have a very positive effect, not only on the learner's listening skills, but it may also help the learner integrate more quickly and more easily into the language community of the target country. It seems that the context for Chinese students studying in Dublin, for example, is precisely that of the community which surrounds them. By using authentic lesson material in the form of a Speech Corpus containing samples of the 'real' spoken language as found within the Dublin community (whether student community or socio-cultural community of the Dublin citizens at large) and the student re-classifying this input against their own frame of reference - which is, at that moment, their life in Dublin - one assumes that one is as close as possible to using authentic materials, authentically. An added benefit of the Speech Corpus is that this type of lesson material is less threatening particularly for Asian students as regards certain culturally induced difficulties such as the 'face' issue, as the learner can use this tool as a stand-alone facility.

Brown and Yule (1983: 21) claim that it is essential to give the learner the opportunity to acquire some of the idiosyncrasies of native natural speech, in order to better fit in to the culture of the target language. It is clear that it should be a teacher's goal to help integrate the learners into the language community of the target country and it seems that, apart from focussing on exposing the learner to 'real' English, one should also help them become 'real' people in their new community; 'warts and all', being able to use the target language in a relaxed and informal manner, with all the linguistic shortcomings and hastily construed half-sentences that belong to native speaker speech. Until the availability of the slowdown facility and speech corpora, authentic NS speech posed specific problems for use in the EFL/ESL classroom, which is now discussed.

## 5 The Problem with Spoken NS English

Non-native speakers of English face many challenges when learning the language. English language scholars and EFL/ESL teachers seek to address these many and varied problems while new theories and models constantly emerge and fade in response to such needs. It seems startling however that in the area of EFL/ESL listening, very little has been done to alter the long-standing current model in operation in most EFL/ESL classrooms worldwide. One of the greatest challenges for non-native learners of English is to perceive, recognise and understand rapid, fluent speech, typical of native speakers. The problem is three-fold:

1) Native English speech features, namely elision, assimilation and weak forms, cause aural and processing problems for EFL/ESL learners
2) Current classroom listening techniques fail to adequately address such problems
3) There is a lack of a comprehensive pedagogical model to train students to 'listen as a native speaker listens'.

### 5.1 Native English Speech Features

Typical native English speech is rapid and fluid, incorporating elision, assimilation and weak forms. While most native English speakers have little or no difficulty in understanding such speech, non-native speakers may feel linguistically overburdened, perplexed and short-changed when communicating with a native speaker. This is due to the fact that in an EFL/ESL listening class, students are regularly exposed to careful speech which closely resembles citation form and which is usually created in a recording studio using actors who generally speak slower and clearer than an average native speaker. This is not an adequate illustration of native English speech however. As a result, EFL/ESL learners are usually ill-equipped to deal with typical native English speech when they encounter it. According to (Brown 1990: 6): 'Students whose education has been largely couched in slowly and deliberately spoken English are often shocked to find, when they enter a context in which native speakers are talking to each other, that they have considerable difficulty in understanding what is being said'.

Instead of viewing streamed speech as 'fast… where familiar words, or groups of words, become unrecognisable to the listener', Cauldwell asserts that we need to acknowledge that it is 'normal' (2003: 3). Brown rightly points out that native speech cannot be described as 'slipshod' or 'careless' as all members of a given group use and understand it (1990: 4).

Brown (1990) makes the argument that native speakers are not totally dependent on the acoustic signal alone to infer the meaning of a message. According to her, native speakers employ top-down processing skills when listening to spoken English, but for various social, psychological and other reasons, non-native speakers have difficulty doing this and are therefore more reliant on bottom-up processing, 'foreign learners are less able to bring to bear 'top down' processing in forming an interpretation, and, hence, are more reliant on 'bottom up' processing (1990: 60). This is because native and non-native speakers process language differently, which means non-native speakers rely more on hearing every word to grasp the meaning. For this reason, EFL/ESL learners should not only be made aware of the salient features of speech, namely stressed elements, but also the more obscure elements, namely elision, assimilation and weak forms. This will enable them to operate without various 'segmental clues' and inform them of which clues will be available and which will not when listening to native speech (ibid).

## 5.2 Current Classroom Listening Techniques

Generally, EFL/ESL teachers use listening activities to serve other language-learning goals, such as discussions and writing tasks, rather than teaching listening skills. This prevents learners from working directly with the recordings and attempting to deal with the blur of streamed speech. Field (2003) observes that many ELT practitioners acknowledge this discrepancy and calls for a change in listening pedagogy. Research reveals that the post-listening phase has the least time devoted to it, as teachers are often more anxious for learners to produce and develop their communicative skills rather than spend time with the actual recordings. However, this is perhaps the most important phase of a listening lesson, as learners should be able to work with the recording in order to access the acoustic blur of natural streamed speech, enabling them to adequately recognise and process native English speech. The post-listening phase should include oral and aural work on sections of notable fast extracts from recordings to improve students' perception skills.

Field states that the most common perceptual cause of understanding breakdown is due to 'lexical segmentation, the identification of words in connected speech' (2003: 327). Listening is more problematic for EFL/ESL learners, as there are only pauses every 12 syllables or so in natural speech, making identification of word boundaries problematic.

Non-prominent occurrences of a word are also very difficult to perceive as they are uttered very quickly and are usually truncated, so that they do not resemble their citation forms, even when they are isolated. Because students are generally trained to recognise words in their citation forms only, they are usually unable to recognise these words as they appear in streamed speech and believe it is a failure on their part as students rather than on the pedagogy, which does not address the issue of streamed speech. Cauldwell asserts that ELT experts, 'need to look at real speech...derive a description, a phonology for listening, that is pedagogically viable' (2004: 10-11).

The transcript of recordings, which is referred to in many EFL/ESL listening classes, does not indicate pertinent features of streamed speech, such as which words will be the most or least blurred or how they will be arranged syntactically. In listening classes, activities relating to the recordings are diverting attention away from the recordings themselves. Field proposes that listening activities should directly relate to listening goals, 'we need to concern ourselves more than we do with speech as a physical phenomenon–with what English sounds like to the non-native listener and with the features which cause obstacles to understanding' (2003: 325). He recognises that the problem with English language listening classes is that not enough time is spent with the actual listening material –'the signal', while 'higher-level understanding' has been the main focus of such listening exercises, 'it encourages us to focus upon the product of listening (in the form of answers to questions) but tells us nothing about the process' (2003: 326). Like Caudwell, Field calls for, 'adequate training in strategies which compensate for gaps in word recognition' (ibid: 325). One needs to observe the actual sounds of English, how they sound to non-native listeners and identify the elements which make comprehensibility and/or understanding problematic for non-native listeners. Some bottom-up processing is necessary before top-down processing can occur. Field notes that, ‘many high-level breakdowns of communication originate in low-level errors’ (ibid). Errors, even small ones such as phoneme discrimination, can affect how a phrase, sentence, even an entire text, is interpreted.

The current EFL/ESL listening class model–mostly comprising carefully recorded material using actors in studios and activities which take the students’ attention from the speech signal to other goals such as speaking or writing–is ineffective in training learners of English to be adept at listening to and processing English speech, particularly that which is typical of native speech. This model needs to be replaced with a pedagogically sound one which enables learners to spend time with the signal, work out the truncated, weak and messy forms of the blur of such speech, in order to be adequately equipped to deal with natural, authentic, native speech, especially English.

### 5.3 Training EFL/ESL Learners in NS-Like Listening Skills

According to Field, the goal of an EFL/ESL listening class should be diagnostic and, ‘provide us with insights onto where understanding has broken down–insights which we can then follow up with small-scale remedial exercises which aim to prevent errors of interpretation…from occurring again’ (2003: 326).

While more natural recordings are being used, particularly at more advanced learning levels, teachers are not informing their students about the features of fast spontaneous speech, ‘we need to be rather more persistent than at present in determining why breakdowns of understanding have occurred’ (Field, 2003: 327). Cauldwell (2003) proposes bridging the gap between what learners hear in the classroom and what they actually experience in the real world by providing EFL/ESL teachers and students with an adequate description of the features of fast spontaneous speech, so that learners are aware of what they have to aspire to in order to be better listeners of English. He believes that, in order to adequately address the needs of learners for listening to natural spoken English, a phonology incorporating the features of streamed speech needs to be developed along with a means for its application, ‘to bridge the gap between slow and fast speech’ (ibid: 2).

Cauldwell (2002b: 5) asserts that EFL/ESL teachers should provide small groups of learners with access to recorded speech acts, which they control—meaning they can re-hear it as often as they need, thereby focussing on their own needs. Getting learners to report on which parts of the recording they found difficult or easy will inform (and may surprise) the EFL/ESL teacher of students’ perception and understanding difficulties. EFL/ESL teachers, through adequate training, should able to observe and explain the features of fast speech and thus teach English language learners how to be better listeners by improving students’ perception and comprehension in a similar way to native speakers, ‘the skill of understanding without attending to every word is a goal to be reached, not a means of getting there’ (ibid: 2). This ‘fast speech phonology’ should include notable features of fast speech, including elision, assimilation, sentence stress and tone units (ibid: 4-5).

Brown points out the importance of slow speech for students in the early stages of learning English, ‘in the early stages, while the student is still struggling with an unfamiliar sound system, not to mention exotic syntactic and lexical forms, this (carefully and slowly enunciated speech) is clearly the only practicable approach’ (1990: 158). She continues by making the point that it is of utmost importance that as students progress, they move beyond careful and slow speech to more natural forms, which will enable them to cope with streamed speech as it is naturally spoken by native English speakers, 'From the point of view of understanding ordinary spoken English, the failure to move beyond the basic elementary pronunciation of spoken English must be regarded as disastrous for any student who wants to be able to cope with a native English situation' (ibid: 158).

This is one of the main applications of the speech corpus, in that students of all levels can control the pace of their listening progress by listening to native English speakers using natural, streamed English while allowing them to slow down the speech (without tonal distortion) if they need to 'catch' unstressed forms, such as elisions and other reductions. Brown (1990: 160) notes that of the many listening materials currently available, the most effective and useful for the foreign student are those that include a wide variety of speech from real situations, by different speakers. This helps to prepare students for when they face native English speakers in the 'real' world.

According to Field, apart from locating lexical boundaries, another problem for non-native listeners is the way words in connected speech are modified so they do not resemble their standard citation forms (2003: 329). Field discusses features such as reduced forms, assimilation and elision and suggests methods for EFL/ESL learners to overcome them, notably, 'to be aware of them, and to be prepared to practise them intensively if there are signs that they are preventing learners from identifying familiar words because of the special conditions of connected speech' (ibid: 332).

The post-listening phase should include oral and aural work on sections of notable fast extracts from recordings to improve students' perception skills. Cauldwell states this is necessary as, 'perception – particularly the ability to hold sounds in short term memory long enough to inspect them for meaning – is a skill that is a prerequisite for understanding' (2002b: 6). For this reason, it is important that students get adequate time with the recordings themselves, that they can re-hear listening passages as often as needed. The focus of the listening task, according to Cauldwell, should directly relate to the central meaning of the recording while also challenging the listeners in terms of perception (ibid).

The next section shows how the theoretical requirements discussed above can be put into practice through the development of a speech corpus, allied to the application of the slow-down facility.

## 6 The Structure of the Speech Corpus

In a traditional corpus used by students of language and applied linguists, size matters. It matters because it is important to provide a representative corpus to permit study of the way language is actually used, rather than to impose prescriptive rules for its use. A good corpus will therefore strive to include as many styles and registers of language as possible, including samples of spoken language.

But in the latter category, what is actually recorded is not the speech itself, but rather the *transcribed* version of spoken language. Normally an attempt is made to remain as faithful as possible to the 'imperfections' of the original speech acts, without giving a phonetic rendition, but the learner or researcher does not have the opportunity to study the original speech itself.

Corpus linguists and students who use corpora to advance their linguistic ability, are generally more interested in the product of speech rather than the speech acts themselves, i.e. speech production. They want to investigate lexical items in context, to better understand the scope of individual words or phrases by studying their collocations and the contexts in which they are used.

Students of pronunciation, on the other hand, and language learners who wish to study the rapid speech of native speakers, have no similar tool to help them, and the language research group at the Dublin Institute of Technology (DIT) have undertaken to develop such a resource. This involves the cooperation of linguists, computer scientists and engineers specialising in the area of digital signal processing (DSP) to produce a tool capable of satisfying the needs of the student of spoken language as well as advanced researchers. Initial work on the corpus has been funded internally. The bulk of the work, however, will be done using funding from the EU *SALERO* project, which commenced in January 2006. DIT is the leader of the workpackage devoted to language synthesis informed by a speech corpus and the work is co-ordinated by the Digital Media Centre.

The three key stages in producing a speech corpus are the recording of spoken material, its transcription and the development of a speech concordancer which links the transcripts with the recordings of the original speech acts. Each stage presents a series of theoretical and technical difficulties and involves the making of strategic choices.

### 6.1 The recording stage

Early on, a decision was made to concentrate on recording dialogues rather than the more easily produced monologues available from broadcast sources—at a price. The bulk of recordings made to date involve a setup where the interviewee is videoed using a digital camera placed behind the interviewer, who wears a lapel microphone. The camera is supplied with a directional microphone. It is the task of the interviewer to engage the interviewee in conversation on a topic of interest to him/her. After a few minutes, the person being recorded relaxes to the extent that s/he engages in genuine conversation and the desired features of native-to-native speech can be recorded. It must be conceded that such conversations cannot be totally natural, since most people find it difficult to ignore the presence of a running camera. However, every effort has been made to elicit relaxed speech from the people being recorded. In every case to date the interviewees were either related to the interviewer or were sufficiently well known to him so as to ensure that genuine conversations ensued after a few minutes. In fact, several sections of the recording were unusable, due to the occurrence of cross-talk when the interviewer introduced backchanneling comments.

### 6.2 Production of an ‘orthotext’

The next critical decision made was to transcribe the recordings in orthographic, or idealised form, which we call the orthotext. This might seem counter-intuitive, given that every effort has been made to ensure that the dialogues recorded represent natural, relaxed speech. On the other hand, it would be very difficult to find the reduced forms characteristic of native speakers in the speech signal unless these were located via the written transcript. The transcript, therefore, is not meant to be a phonetically accurate representation of the dialogues, but rather provides a search mechanism to locate the speech acts which are of interest to the user. It remains true to the semantic content of the dialogues recorded, but has been normalised so as to ensure that all occurrences of the search strings can be found and contrasted. This transcription stratagem should not be construed as a weakness, but rather as a focussed mechanism for finding the relevant section of the speech signal. It is also conceivable that outputs from the *SALERO* project might be used to synthesise speech based on the normalised transcription, which provides a further justification for adopting the current approach.

A further use of the orthotext transcript is as a guide to the learner, in that it represents a normalised or idealised form of the language likely to have been acquired by learners in their basic course materials. By this means, users can therefore contrast the actual speech signal with the normalised form represented by the orthotext, and where notable discrepancies occur, the relevant section can be slowed down to allow the learner to follow the phonetic stream actually produced by the native speaker. While this mechanism will not provide a closely aligned link between NS speech (especially weak/reduced forms) and the orthotext, it will be of benefit to the learner to follow the perceived speech production of the NS speaker slowed down, in the same manner as it is of assistance to the novice tennis player to see a slowed-down version of a professional tennis serve, using fast photographic techniques. Here too speed and efficiency are achieved at a pace too fast for the eye to follow. In a similar fashion it is hoped that allowing the ear to follow–in slow motion–the articulatory performance of the NS speaker will enable learners to hear and practise the articulatory gestures required to imitate the original recording, if it is desired to sound more like a NS using that particular language model. At the very least, it should be possible to enhance learners' listening skills by imitating NS production, using the slow-down facility.

The exact nature of orthotext will be a matter for experimentation and debate over the coming months, but for the moment it is simply a written record which normalises the spoken text to complete, citation-form words. Thus 'it's' is rendered as 'it is' and 'whydje do that' as 'why did you do that'. It would be difficult, if not impossible, to locate the reduced form 'dje'–representing 'do you' or 'did you'–in the original sound file unless these were recorded in full lexical form in the transcript. If native speakers (NSs) were not understood at the first pass of a speech act and asked to repeat it in a 'second pass' utterance, then they will accommodate the listener by making a greater effort to be more intelligible. It is likely that the speech will remain relaxed and conversational, but it will be clearer. It will, however, have the same semantic content; it is simply that the listener's needs have been prioritised over the convenience of the speaker in communicating with the least effort.

This is almost exactly analogous to what happens in handwriting. A note scribbled to an intimate acquaintance might well be indecipherable by someone unknown to the writer. If the latter is aware that a stranger will read the note, then greater care will be taken to make the handwriting legible. It will still be individualised handwriting, but the reader of the note is more likely to be able to make out individual letters. The following illustration of a signature and its increasing clarity will exemplify the process.

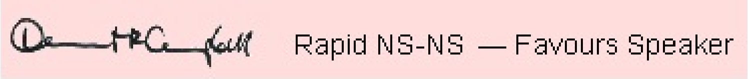


**Figure 1**

In a similar fashion to phonemes in the stream of NS speech, individual letters cannot be recognised in the above signature. A signature's function is to have a unique shape, not to make its constituent components individually recognisable.

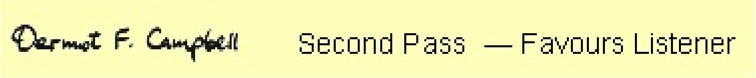


**Figure 2**

In the above version an attempt was made to facilitate the reader in identifying the script. Here, however, the 'register' of the communication has changed and it is no longer a signature, but rather the hand-written, orthographic version of a signature.

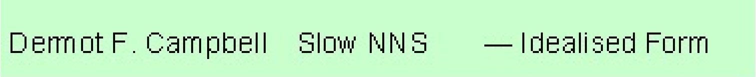

Dermot F. Campbell
Slow NNS — Idealised Form

**Figure 3**

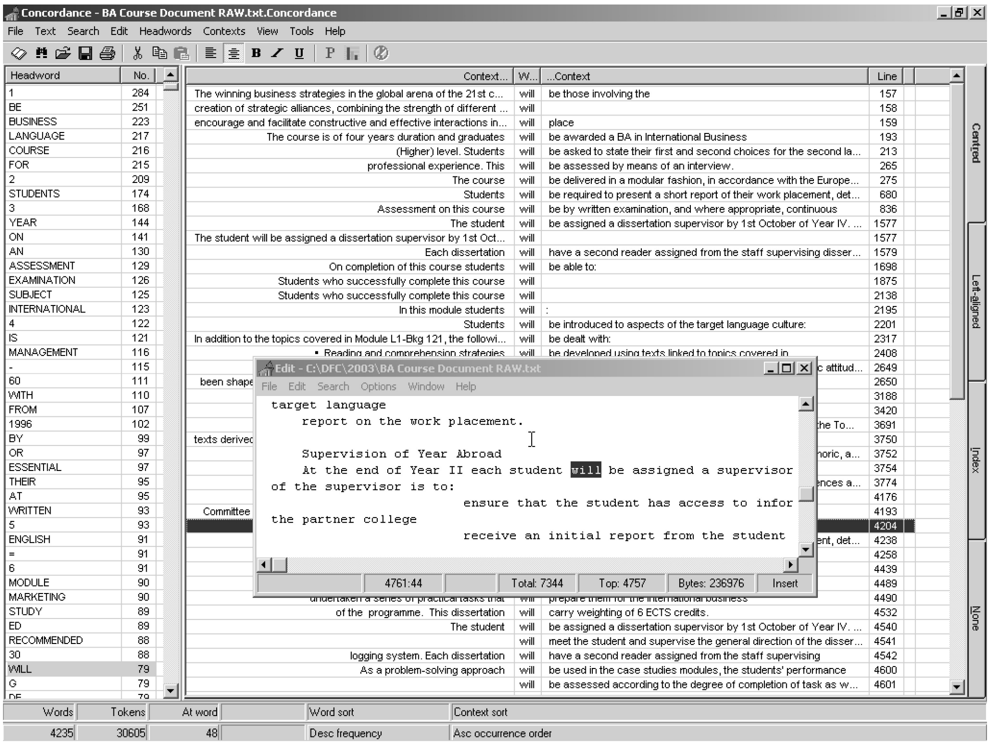
The printed version of the name (or its equivalent in block capitals) corresponds more accurately to the idealised form of the written word familiar to the language learner. Almost nobody writes like that, but this is the form of the language that learners tend to internalise. It is the gap between the first and the third versions of the communication that the speech corpus hopes to help bridge.

### 6.3 The concordancer links speech and orthotext

The third important stage in developing the speech concordancer is its division into meaningful units while maintaining the unity of the overall recording.

Each transcript will be broken down into 'timed units'– i.e. stretches of speech which were dictated by the speaker, whether to afford thinking time or to plan the next speech segment – and these in turn will be linked to the relevant section of the soundwave. Initially this will be done manually, but effort will be devoted to (semi-) automating this process, which is labour intensive. As recognised at *EuroCALL 2005*, however, the time has come for quality in a corpus to take precedence over quantity. Since the primary requirement of the orthotext is to locate speech acts relevant to the study of spoken language, it is anticipated that there will be significant differences in the architecture of a traditional concordancer and a speech concordancer.

The diagram below illustrates a conventional concordancer which allows a search string to be located in a body of running text, and for each occurrence to be displayed line by line, centred on the search string itself. Clicking on a line will cause a window to pop up showing the wider context in which the display line is embedded. These features are effective in tracking collocations and in scoping the semantic fields of the search criteria, but have little relevance in the current context of a speech concordancer.


Concordance - BA Course Document RAW.txt.Concordance
File Text Search Edit Headwords Contexts View Tools Help
Centred
Left-aligned
Index
None

| Headword | No. |
|---|---|
| 1 | 284 |
| BE | 251 |
| BUSINESS | 223 |
| LANGUAGE | 217 |
| COURSE | 216 |
| FOR | 215 |
| 2 | 209 |
| STUDENTS | 174 |
| 3 | 168 |
| YEAR | 144 |
| ON | 141 |
| AN | 130 |
| ASSESSMENT | 129 |
| EXAMINATION | 126 |
| SUBJECT | 125 |
| INTERNATIONAL | 123 |
| 4 | 122 |
| IS | 121 |
| MANAGEMENT | 116 |
| - | 115 |
| 60 | 111 |
| WITH | 110 |
| FROM | 107 |
| 1996 | 102 |
| BY | 99 |
| OR | 97 |
| ESSENTIAL | 97 |
| THEIR | 95 |
| AT | 95 |
| WRITTEN | 93 |
| 5 | 93 |
| ENGLISH | 91 |
| = | 91 |
| 6 | 91 |
| MODULE | 90 |
| MARKETING | 90 |
| STUDY | 89 |
| ED | 89 |
| RECOMMENDED | 88 |
| 30 | 88 |
| WILL | 79 |
| G | 79 |
| DE | 79 |

| Context... | W... | ...Context | Line |
|---|---|---|---|
| The winning business strategies in the global arena of the 21st c... | will | be those involving the | 157 |
| creation of strategic alliances, combining the strength of different ... | will | | 158 |
| encourage and facilitate constructive and effective interactions in... | will | place | 159 |
| The course is of four years duration and graduates | will | be awarded a BA in International Business | 193 |
| (Higher) level. Students | will | be asked to state their first and second choices for the second la... | 213 |
| professional experience. This | will | be assessed by means of an interview. | 265 |
| The course | will | be delivered in a modular fashion, in accordance with the Europe... | 275 |
| Students | will | be required to present a short report of their work placement, det... | 680 |
| Assessment on this course | will | be by written examination, and where appropriate, continuous | 836 |
| The student | will | be assigned a dissertation supervisor by 1st October of Year IV. ... | 1577 |
| The student will be assigned a dissertation supervisor by 1st Oct... | will | | 1577 |
| Each dissertation | will | have a second reader assigned from the staff supervising disser... | 1579 |
| On completion of this course students | will | be able to: | 1698 |
| Students who successfully complete this course | will | | 1875 |
| Students who successfully complete this course | will | | 2138 |
| In this module students | will | : | 2195 |
| Students | will | be introduced to aspects of the target language culture: | 2201 |
| In addition to the topics covered in Module L1-Bkg 121, the followi... | will | be dealt with: | 2317 |
| ▪ Reading and comprehension strategies | will | be developed using texts linked to topics covered in | 2408 |
| | | c attitud... | 2649 |
| been shape | | | 2650 |
| | | | 3188 |
| | | | 3420 |
| | | the To... | 3691 |
| texts derivec | | | 3750 |
| | | horic, a... | 3752 |
| | | | 3754 |
| | | nces a... | 3774 |
| | | | 4176 |
| Committee | | | 4193 |
| | | | 4204 |
| | | ent, det... | 4238 |
| | | | 4258 |
| | | | 4439 |
| | | | 4489 |
| undertaken a series of practical tasks that | will | prepare them for the international business | 4490 |
| of the programme. This dissertation | will | carry weighting of 6 ECTS credits. | 4532 |
| The student | will | be assigned a dissertation supervisor by 1st October of Year IV. ... | 4540 |
| | will | meet the student and supervise the general direction of the disser... | 4541 |
| logging system. Each dissertation | will | have a second reader assigned from the staff supervising | 4542 |
| As a problem-solving approach | will | be used in the case studies modules, the students' performance | 4600 |
| | will | be assessed according to the degree of completion of task as w... | 4601 |

Edit - C:\DFC\2003\BA Course Document RAW.txt

File Edit Search Options Window Help

```
target language
    report on the work placement.

    Supervision of Year Abroad
    At the end of Year II each student will be assigned a supervisor
of the supervisor is to:
                    ensure that the student has access to infor
the partner college
                    receive an initial report from the student
```

| | | | | | | |
|---|---|---|---|---|---|---|
| | 4761:44 | | Total: 7344 | Top: 4757 | Bytes: 236976 | Insert |

| Words | Tokens | At word | | Word sort | Context sort |
|---|---|---|---|---|---|
| 4235 | 30605 | 48 | | Desc frequency | Asc occurrence order |

**Figure 4**

In contrast, what is of importance in a speech concordancer is the immediate phonetic environment in which lexical and phrase units are spoken. It is envisaged that each phrase will be displayed as in the figure above, but with the difference that clicking on each line (i.e. timed unit) will play the relevant original recorded segment rather than reveal co-text. Each occurrence of 'did', for example will appear in a separate line in order to enable comparison, and it will be possible to compare the different ways in which the lexical item 'did' is pronounced, ranging from a hyper-articulated version to the most severely reduced form compatible with intelligibility by a NS.

## 7 Speed: Fast and slow

The architecture of the speech corpus envisages the inclusion of a speed index which will indicate the rate at which an individual timed unit was delivered. This will be of assistance in indicating the likelihood of finding reduced speech forms in the corpus—a feature of rapid native-to-native speech which causes considerable difficulty for learners of the language in question. Having entered the search string in order to locate the various instances of the string contained in the corpus, the displayed list will be capable of being further sorted according to rate of delivery. Initially this rate of delivery will refer to the speed at which the timed unit as a whole was spoken, but will still give a fair indication as to the likelihood of the presence of reduced or weak spoken forms. Whether further segmentation to word level is possible or economical remains to be discovered.

As can be seen from the table below, a reduced form such as 'Whydjedothat?', which took 863 ms to deliver, has the full semantic content–for a native listener–of its second-pass or (here) citation form equivalent: 'Why did you do that?'. The orthotext message was spoken at 278 syllables per minute, whereas the speed of delivery of the same question in reduced form was 348 syll/min, i.e. in other words the same message was conveyed in just under 80% of the time. Were both waveforms contained in the audio corpus and transcribed identically, but with their speed index included, then clearly a search for the word ‘did’ would juxtapose both orthotext versions. When further sorted by speed index–which will be visible in the concordanced transcript–this will be a clear indication that the faster version exhibits reduction characteristics. This supposition can then be verified by clicking alternately on both transcription lines to play and contrast both recordings.

| Spoken | Secs. | Syll. | Orthotext | Syll. | Syll./Min |
|---|---|---|---|---|---|
| Whydjedothat? | 0.863 | 4 | Why did you do that? | **5** | **348**<br>(278) |
| I’ll be 5 mints. | 0.986 | 5 | I will be 5 minutes. | **6** | **365**<br>(304) |
| ’Mon quick! | 0.497 | 2 | Come on quick! | **3** | **362**<br>(241) |

**Table 1**

### 7.1 Initial Findings–Chunks

Several authors such as Fillmore (1979), Klatzky (1980) and McCarthy and Carter (2002) have drawn attention to the existence and function of chunks in natural dialogue. There is evidence in the transcripts studied to date that chunking is very much linked to speaking rate (see figure below). The interviewer, Marc, has an average speed of delivery in the interview with Aelish (his mother) of 388 syllables per minute. He makes frequent use of the chunk ‘That’s right’, which has been transcribed as ‘that is right’. The speeds recorded for this chunk in the recording are: 741, 720, 677, 627, 411, 371 and 289 syllables per minute.

At speeds considerably above his average speaking speed, Marc’s chunk *that’s right* has the characteristics of a throw-away remark meant to provide encouraging feedback. It displays a lack of emotional engagement and serves mainly to mark attention to the conversation. At the slower speeds, however, the chunk takes on a more pragmatic character and testifies to a more personal involvement and active recalling of the events spoken of. The unity of the above-average-speed chunk would appear to dissolve, the lexical independence of the constituent parts assert itself and a pragmatic element creep into the use of the chunk.

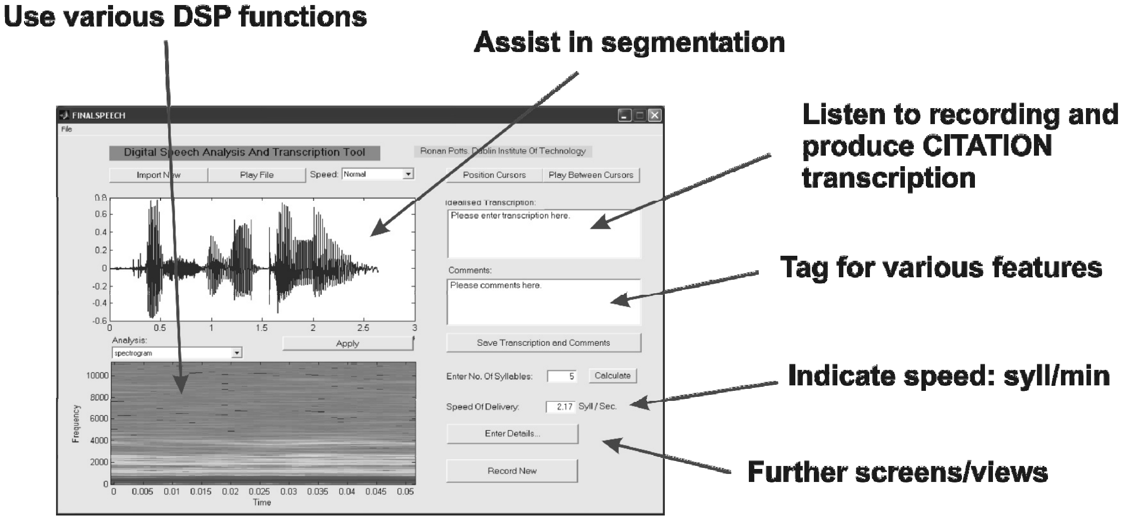
More work will have to be done on this phenomenon, but early indications are that speed-indexing might shed an interesting light on the distinction between the terms formulaic sequence, as used by Wray (2002, 2004, 2005) and chunks proper. At first sight, it would seem that while all chunks are also formulaic sequences, not all formulaic sequences are chunks, in the sense that these seem to be stored, recalled and articulated as units. In order to study this phenomenon further, it will be necessary to study a wider range of samples.

In any event, however, there is a clear relevance to CALL in that the learner (or researcher) will have access to original recordings of chunks/formulaic language which can be sorted according to a useful speed of delivery index (given in syllables per minute) and then compared, so that pragmatic differences can be highlighted. It is hoped to allow concatenation of timed units in such a way as to display the wider phonetic context in which the chunks were spoken, since the timed units themselves are generally too short to highlight the contrast between chunks and their more lexically independent environment.

## 8 Transcription tool

A further functionality of the planned speech corpus will be the ability to slow down the original WAV file to any desired speed without tonal distortion, using the *DITCall* algorithm developed by the Digital Media Centre. This will allow students and researchers the ability to study the original speech acts proper rather than simply examine the semantic import of the original communication via the transcript. This in turn means that NS recordings can be used in class, and by individuals who want their oral performance in the language studied to be more native-like. It can also be used by all students of the spoken word who want to develop a better facility in understanding native-native exchanges. It is anticipated that the fully developed speech corpus will meet Brown and Yule's sociolinguistic requirements (1983: 21) and help to avoid exclusion from full participation in the target language community.

At the moment the corpus is being constructed by hand and the video recordings transcribed and segmented manually. This is such a labour-intensive activity that ways will be sought to speed up the process. One such tool currently in development is the Transcription Graphical User Interface (GUI) below, which can be used to aid the tagging and transcription.



Use various DSP functions

Assist in segmentation

Listen to recording and produce CITATION transcription

Tag for various features

Indicate speed: syll/min

Further screens/views

Speed: Normal

Idealised Transcription: Please enter transcription here.

Comments: Please comments here.

Analysis: spectrogram

Enter No. Of Syllables: 5

Speed Of Delivery: 2.17 Syll / Sec.

**Figure 5**

The Transcription GUI acts as an interface between the text which will form the transcription corpus and the original WAV file. Via the GUI the transcriber can load, display and play the original sound file. There will be sufficient digital signal processing (DSP) functionality available to select and zoom in to the wave so that section boundaries can more easily be distinguished. The duration of the section chosen will also be displayed, so that speed of delivery can be automatically calculated once the syllable count has been established. This is a further area where natural language processing tools might assist in automation of the process.

It is planned that the Transcription GUI should be capable of facilitating the tagging of the transcribed text in a manner compatible with the goals of the EU *SALERO* project. This project will study the possibilities of automatically re-purposing digital objects by building meta-descriptions into the initial creations. One such digital object is, of course, digitally recorded speech, and the developing speech corpus will be a vital tool in realising this ambitious goal. While *SALERO* is concerned primarily with reusable digital objects, its rule-based algorithms will be derived from the same speech corpus described in this article from the perspective of the human listener.

## References

**Bacon, S., & Finnemann, M.** (1990). A study of the attitudes, motives, and strategies of university foreign language students and their disposition to authentic oral and written input. *Modern Language Journal, 74*, 459-473.

**Brown, G. and Yule, G.** (1983) *Teaching the spoken language. An approach based on the analysis of conversational English.* New York: Cambridge University Press. (p.21)

**Brown, G.** (1990) *Listening to Spoken English*, London: Longman.

**Carter, R.** (2003) Teaching about talk – what do pupils need to know about spoken language and the important ways in which talk differs from writing? In: New Perspectives on Spoken English in the Classroom. Discussion Papers. London: Qualification and Curriculum Authority

**Cauldwell, R.** (2002a) Phonology for listening: relishing the messy, richard@speechinaction.com (accessed April 08, 2004).

**Cauldwell, R.** (2002b) Grasping the nettle: the importance of perception work in listening comprehension, http://www.developingteachers.com/articles_tchtraining/perception1_richard.htm (accessed April 12, 2004).

**Cauldwell, R.** (2003), Streaming Speech: Listening and Pronunciation for Advanced Learners of English, Birmingham: speechinaction.

**Cauldwell, R.** (2004), 'Stuck in TAR: how we prevent learners from handling everyday speech', Speak Out! – newsletter of IATEFL Pronunciation Special Interest Group, 32: 8-11.

**Cook, G.** (1998) The issues of reality: A reply to Ronald Carter. *ELT Journal, Volume 52/1 January 1998.* Oxford University Press

**Crystal, D.** (1981) *Directions in Applied Linguistics.* London: Academic Press. (p.90-92)

**Darian, S.** 2001. Adapting Authentic Materials for Language Teaching. *Forum, Vol. 39, No 2, April-June 2001.* (p.2) Bureau of Educational and Cultural Affairs, US Department of State

**Fillmore, C.J.** (1979). On Fluency. In: Individual differences in language ability and language behavior. C.J. Fillmore, D. Kempler & S.-Y.W. Wang (eds.) New York: Academic Press (pp.85-101).

**Field, J.** (2003) Promoting perception: lexical segmentation in L2 listening, ELT Journal, Vol 57/4 Oct 2003, Oxford: OUP.

**Herron, C., & Seay, I.** (1991). The effect of authentic aural texts on student listening comprehension in the foreign language classroom, *Foreign Language Annals 24,* 487-495.

**Kang, S.** (1997) Factors to Consider. Developing Adult EFL Student's Speaking Abilities, Forum, Vol. 35 No 3, July-September 1997. (p.8) Bureau of Educational and Cultural Affairs, US Department of State

**Klatzky, R.** 1980. Human memory. New York: W.H. Freeman.

**McCarthy, M. and Carter, R.** (2002) This that and the other: Multi-word clusters in spoken English as visible patterns of interaction. Teanga (Yearbook of the Irish Association for Applied Linguistics), vol. 21, (pp. 30-52).

**Mishan, F.** (2004) Authenticating corpora for language learning: a problem and its solution, *ELT Journal Volume 58, 2 July 2004*. Oxford University Press

**Tribble, C.** 1997. Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching. In: PALC '97 Proceedings, J. Melia and B. Lewandowska-Tomaszczyk (ed.) Lodz: Lodz University Press

**Widdowson, H.G. (**1980) Models and Fictions, *Applied Linguistics, Vol. 1, No. 2, summer 1980.* Oxford University Press. (pp. 165-170)

**Widdowson, H.G.** (2000) On the limitations of linguistics applied, *Applied Linguistics* 21/I: 3-25

**Wray, A.** 2002. Formulaic Language and the Lexicon. Cambridge: Cambridge University Press (p. 9)

**Wray, A.** 2004. Formulaic language learning on television. In: Formulaic Sequences, N. Schmitt, (Ed.). Amsterdam: John Benjamins Publishing Company (pp.249-268)

**Wray, A.** 2005. 'Idiomaticity in an L2: linguistic processing as a predictor of success.' Keynote address at IATEFL Conference, Cardiff, 2005