McCarthy, M.J. and **O'Keeffe, A.** (2012) "Analysing Speech Corpora". In T. Cobb (Ed.) *The Encyclopedia of Applied Linguistics.* New York: Wiley-Blackwell, pp.104-112.

Analyzing spoken corpora

Michael McCarthy
University of Nottingham
michael.mccarthy@nottingham.ac.uk
Anne O'Keeffe
Mary Immaculate College, University of Limerick
Anne.OKeeffe@mic.ul.ie
Word count: 2752
Reference word count: 702

Key words corpus spoken analyzing tools transcripts concordances chunks

Spoken corpora are collections of spoken texts, either in the form of transcripts alone or accompanied by audio- or audiovisual recordings, stored in a computer and made available for analysis using customised software programs. Spoken corpora have grown in importance in recent years and a number of major corpora now exist for English and other widely-used world languages. Spoken corpora are typically compiled using data from a range of sources, sometimes by very general demographic sampling, or sometimes in more specialized contexts. Transcription of the data may be broad or narrow, but usually includes some kind of non-verbal or non-linguistic data. The tools of corpus linguistics can be applied to spoken data in a variety of ways, enabling researchers to learn about relative frequencies of words and patterns in different types of spoken data or in contrast with written data.

[A]Introduction

Spoken corpora are collections of transcripts of real speech, typically consisting of speech events recorded on analogue or digital recording equipment and then transcribed and stored in a computer as text files. *Spoken corpora* are often distinguished from *speech corpora*: speech corpora are usually collections of speech (which could be anything from transcripts of spontaneous speech to recordings of people reading out loud prepared lists of single words) that are compiled for purposes such as creating automatic voice-to-text applications, telephone technology or the analysis of the phonetic substance of speech (Harrington, 2010). Researchers who create speech corpora are not necessarily concerned with what is said; they are more concerned with how it is said, i.e. the speech signal itself. Researchers who work with spoken corpora are typically interested in what people say, why they say

it and how they use spoken language to communicate their messages and to interact with one another. Spoken corpora are, by definition, a recent development; before tape recorders became common, the only way to record people's natural spoken interactions was by observation (or eavesdropping) and attempting to write down what was said (e.g. Timmis, 2009).

[A]Characteristics of spoken corpora

Spoken corpora may be of different kinds and may include anything from formal, monolog events such as lectures and speeches to informal conversations between two or more parties. The settings where recordings are made may vary from people's homes and workplaces to stores, restaurants, broadcasting studios, and classrooms. Another dimension along which spoken corpora may vary is the purpose of the interactions they contain: the conversations might be debates, transactions of information, goods and services, narratives, media interviews, therapeutic sessions, etc. (see McCarthy, 1998 for a description of the various contextual settings that went into the construction of one spoken corpus). Finally, the participants who agree to be recorded for the corpus may be a cross-section of the general population (see, for example, information on the demographic sampling that underpinned the spoken segment of the British National Corpus at http://www.natcorp.ox.ac.uk/corpus/creating.xml), a particular social group, for example, teenagers (as exemplified in the COLT corpus of London teenage speech; see Stenström *et al* 2002), or a specialised professional group, for example, business people (Handford, 2010). A spoken corpus will usually have speaker information sheets linked to each recording which allows for socio-demographic research about language use across age, gender, geographical origin of the speaker, and so on (see Crowdy, 1993).

[A]Understanding the spoken transcript

Before a corpus is analyzed, the compilers make decisions about how to transcribe it. Consider this extract from the Limerick Corpus of Irish English (see Farr, Murphy, and O'Keeffe, 2002 for details of the corpus).

<$2> I came up with all kinds of things+

<$1> Umhum.

<$2> +I found that even though I wasn't able to look at my tape I realised that I hadn't eh my instructions were not clear at the+

<$1> Umhum.

<$2> +because they weren't sure exactly what I wanted I didn't check and see to ch= see if they really understood the probability factor and <$02> <$G?> <\$02>+

<$1> <$02> Yes <\$02>.

<$2> + <$G?> and I just kind of went bluh here it is.

<$1> Umhum.

<$2> So it's like I just threw it at <$X> 'em | them <\$X> and I feel badly about that em I let the students distract me with their comments and their ideas.

Everything marked with unfamiliar symbols (such as the diamond brackets < >, the dollar signs, the backslashes, etc.) is there to assist the analyst in some way. The tags <$1> and <$2> mark the individual speakers, numbered in the order in which they first speak. Somewhere in an accompanying database, we can find out their sex, age, social background, where and when the conversation occurred, and so on, and, if necessary, we will find a file number than can take us back to the audio recording. With the latest digital technology, we may also be able to play the audio recording while we read the transcript. The plus signs (+) indicate that someone has not finished what they were saying and that someone else has said something in the middle of the utterance (e.g. *umhum*), which then continues. Symbols such as <$G?> indicate that the transcriber could not make out what the speaker was saying. The backslash \ indicates the end of a numbered overlap, where two speakers speak simultaneously. The vertical line | separates a non-standard form (in this case *'em* from the standard form *them*). All of these conventions are invaluable to the analyst, for reasons we shall see below. In addition, transcribers have to decide how to transcribe vocalizations that are not usually recognised in the dictionary as 'words' (e.g. *umhum*, *bluh, eh*), or half-uttered words such as *ch*=. Other decisions may include whether to transcribe the corpus showing intonation and stress patterns (Cheng and Warren, 2000; Cheng, Greaves, and Warren 2005; Warren, 2004).

The presence or absence of extra levels of detail in a transcript is usually dictated by available time and resources: transcription is time-consuming and expensive, and researchers often have to economise and omit features in the transcript which might have been invaluable to the analyst. Whatever the resource limitations, transcribing always involves choices, and no-one can transcribe absolutely every relevant feature of a spoken interaction (Cook, 1990). What matters to the analyst is to have important features transcribed so that they are automatically searchable wherever possible.

Increasingly, spoken corpora are supplemented by linked video images, which offer further analytical support to the written transcript and audio material (Adolphs and Knight, 2010; Thompson, 2010).

[A]Analytical tools

Different analytical techniques will provide a range of results which reveal distinct aspects of the spoken material. With free or proprietary software suites, it is usually possible to generate frequency lists which arrange all the words in the corpus in order of frequency, from the most common to the rarest (see Scott 2010; Evison 2010). With such lists, we can observe important differences in the distribution of words between, for example, speech and writing, or between different types of speech (e.g. monolog versus dialog). In the British National Corpus (BNC - a collection of 100 million words of data, of which 10 million are spoken), the noun *kids* is twice as frequent in the spoken segment of the corpus as in the written fiction segment, and almost 20 times more frequent than in the written academic segment. The noun *children* is much more evenly distributed across the three datasets. Conversely, a frequency list can tell us what words are rare in spoken language; for example, the words *vicarious* and *simulacrum*, which occur in academic writing, occur nowhere in the spoken BNC. For further information on the different distributions of words in the BNC, see Leech et al (2001).

It is not only single words that can be analyzed in a corpus, and good software will reveal that spoken language manifests a huge number of ready-made 'chunks' (strings of two or more words sometimes referred to as *n-grams, lexical bundles, lexical phrases, clusters, multi-word units*) which speakers use repeatedly, enabling fast retrieval of items from the mental lexicon and making possible the steady flow of language we associate with fluency (Biber et al., 1999; Carter and McCarthy, 2006; O'Keeffe et al., 2007; Cheng et al., 2009; McCarthy, 2010; Greaves and Warren, 2010).The analyst can usually choose the length of the chunks he/she wants the software to retrieve from the data (e.g. all the repeated three-word chunks, or four-word chunks). O'Keeffe, McCarthy and Carter (2007) give lists of common chunks and idiomatic expressions in British and North American English spoken data. These lists reveal that the most common chunks (such as *you know, I mean, that kind of thing*) are highly interactive and paint a clear picture of how speakers take one another into account and work hard to create satisfactory personal relations (see McCarthy and Carter, 2002; see also Shin and Nation, 2008, Martinez and Schmitt, under review). The spoken segment of the American National Corpus (First Release) includes in its list of most common three-word chunks *was nice talking, I understand that, all that stuff,* and *what would you*.

Other analytical instruments based on word-frequency lists include collocation tools which gauge the likelihood of two words occurring together and then measure the actual co-occurrence and

give to the collocation a statistical score of significance. In the case of the spoken language, this may again reveal interesting contrasts both between speech and writing and between different forms of spoken language. In the BNC, for example, the most immediate noun-collocates of the adjective *difficult* in the written academic segment are *question, task, problem,* and *concept*, while *question, situation, job,* and *time* dominate the spoken list, revealing the different preoccupations and priorities that occur in speech as compared with academic writing. Another useful technique is to examine keywords, that is to say words which occur with unusual frequency in one corpus or another. McCarthy and Handford (2004) show that the verb *need* is a keyword in spoken business English, and trace its unusually high frequency to its common use in the chunk *we need to*, denoting corporate exigencies and indicating a way of hedging directives by expressing them as collective requirements (i.e. *we need to* said by a senior or powerful person in an organization may be interpreted by subordinates as *you must/should*). By contrast, *can* is the highest ranking modal verb among the top 20 keywords in a corpus of spoken academic English collected in universities in Ireland (LI-BEL, see Walsh and O'Keeffe forthcoming). This is attributed to its instructional use in this pedagogical context, for example, *Can you tell me, Can you think of.*

Perhaps the most useful and revealing tool for any corpus linguist is the concordance, a computer screen display or printout of a chosen word or phrase in use in numerous different contexts by numerous different users. For spoken language, the concordance can bring together the utterances of many different speakers over many different places and occasions of use, giving to the researcher a powerful instrument for gauging what is typical usage and in what contexts things are said. Most concordance software allows a variable range of context to be viewed, from a single line of text across the screen with the key word(s) centered vertically, to a longer context such as a whole speaker turn. For instance, when exploring concordance lines for the keyword *can* referred to above in the spoken academic corpus LI-BEL, the pattern *can* + *actually* becomes instantly visible once the words that occur after the search word, *can,* are 'right-sorted' alphabetically, that is to say all the words beginning with *a-* (such as *actually*) are shown first, then *b-*, and so on (figure 1). The researcher can then identify, by going back to the source files for extended context, that *can* + *actually* is associated with the marking of new information by lecturers.

Figure 1: Extract from concordance lines for *can* in the LI-BEL academic corpus

And you **can actually** do it you **can actually** do it through what should I say amm
have a history of tuberculosis you **can actually** be refused entry into other states. It's
have it out of synch then maybe we **can actually** build in that delay as well. So that both
ures it can change the figure. You **can actually** change the graph representation to see
people at this point and time. You **can actually** click there on layout. Right? The third
actual ahh P C B. And sometimes we **can actually** consider the actual design of the actual

tion to overload. So here we **can actually** consider aspects like that in that we're
can do an actual performance where **can actually** create the actual workplace and get
mechanisms for the student they **can actually** decide whether or not they need to do a
hree perform. You **can actually** you **can actually** do it that way. Okay? Now let's just say

[A]Exploiting distinctive features of spoken corpora

Because written data used in the service of writing dictionaries for so long dominated the study of corpora, linguists tended to focus on words, phrases and grammatical phenomena in texts. However, as we have already noted in section 3, above, a transcript may contain a wealth of non-word information (the tags and other conventions we have already looked at) which the spoken corpus analyst can exploit better to understand how common events such as conversations are formulated. We noted that individual speakers were usually designated by a symbol (e.g. a dollar sign with a speaker number). Computer software is impartial to words or non-words and can be instructed either to ignore such speaker tags (if we are only interested in the words uttered) or to search for them and count them, just like words. This latter option enables researchers to learn a lot about how speakers construct their turns at talk, and can offer powerful quantitative underpinning to the micro-analyses of individual conversations provided by researchers working in the conversation analysis (CA) tradition. Simple statistical operations such as counting how many turns take place in a conversation, or what the average number of words per turn is, can provide important insights into particular types of conversations or to talk as a whole. Tao (2003) takes the exploitation of speaker tags further, and provides a frequency list of turn-openers, the first words after speaker tags, as evidenced in a corpus of spoken American English. The words which characteristically open turns are freestanding items such as *yes, so, right, well*, and so on, and they show us that turn-openers predominantly attend to what the speaker has just heard, providing a linking function and contributing to confluence by creating smooth transitions from one turn to the next.

Figure 2 shows part of a concordance for a speaker tag <$*> + *well* (where * is a 'wildcard' representing any character) in the five-million word CANCODE[1] spoken corpus. The data has been right-sorted.

Figure 2: Part of right-sorted concordance for *well* (CANCODE)

at rather than me. <$E> laughter </$E> <$1> Well actually+ <$?F> <$G?> <$1> +<$=> they d
. <$2> Y= is that you're advice <$G2>? <$1> Well actually in in my case it's one of the few wo

[1] CANCODE means Cambridge and Nottingham Corpus of Discourse in English. Cambridge University Press is the sole copyright holder. For details of the corpus, see McCarthy (1998).

> <$=> Because <\$=> Have you been there? <$1> Well actually it was one of the places I was think
to read this very carefully. <$2> Yes. <$1> Well actually no I don't mean that to sound <$E> l
<$?F> How how many people live in <$G?>? <$1> Well actually that that's that's not quite erm the
n people used to leave early didn't they. <$1> Well actually <$=> w= you <\$=> in those days you
> or not. <$1> Er <$G?> <$?F> <$G?> <$1> Well actually <$=> we'll look </$=> we'll look at
``Oh it's reserved'' and that's it. Easy. <$2> Well actually I could have called one of <$X> t' |
shower that you put in five years ago." <$2> Well actually I decided not to use it. <$1> Yea
at? Or do you prefer this sort of format? <$2> Well actually I don't s= I don't suppose No I thin
t <$G?>. So it must be a really big pond. <$2> Well actually if you think about it because it's n
he window. I just can never get it <$G?>. <$2> Well actually <$=> I'm I'm <\$=> I'm not too bad t
't usually be that cold. <$1> No true. <$2> Well actually it might <$G?>. <$1> Mm. Uh-huh.
sorry about that. <$E> laughter <\$E> <$2> Well actually one one of the questions that we did
impress people. <$1> Oh right. I see. <$2> Well actually that sort of stuff does work in some

We see here from the tags immediately before it that *well* frequently begins a speaker's turn, and that *well actually* is a frequent chunk: we do not find *actually well* occurring anywhere in the concordance. Some software allows case-sensitive searching, which would enable us to distinguish *Well* from *well*, making it easier to find turn-opening examples if the transcribers followed the convention of using a capital letter at the start of a turn. With most software too, by clicking on any line in the concordance, we can get back at once to the whole text of the conversation to access more context and work out the situations in which people typically respond with *Well actually*.

Another example comes from a sub-corpus of grocery shop recordings from the LCIE corpus, where the high frequency discourse marker *now* is frequently preceded by the tags <$E> and <\$E>, used to add extralinguistic information (figure 3). The extralinguistic information in all cases is the sound of the cash register as money is handed over, annotated as <$E> *sound of till* <\$E>. This allows us to see the contextual pattern whereby the shop attendant rings up the price of the customer's item on the till, and the attendant announces the price of the item to the customer, always preceded by the discourse marker *now.*

Figure 3: Extract from concordance lines for *now* in shop recordings from LCIE

<$E> sound of till <\$E> **Now** two fourteen thanks. <$1> Two fourteen so
<$E> sound of till <\$E> **Now** two fifteen so please <$E> pause <\$E>. <$E> sound of
<$E> sound of till <\$E> **Now** two sixty seven so please. <$2> Now I have the sixty
<$E> sound of till <\$E> **Now** three twenty so please <$E> sound of coins <\$E>. <$2>
<$E> sound of till <\$E> **Now** sixty eight please <$E> pause <\$E>. Thanks. <$2> I don't

<$E> sound of till <\$E> **Now** one twenty please. Thanks. <$2> Thank you. <$1>
<$E> sound of till <\$E> **Now** eight twenty eight so please. <$E> sound of plastic bags
<$E> sound of till <\$E> **Now** and three nineteen okay. <$1> Thank you. <$2> Thank
<$E> sound of till <\$E> **Now** two forty please. <$E> sound of coins <\$E> Thank you very
<$E> sound of till <\$E> **Now** a pound please. <$1> <$E> sound of till
<$E> sound of till <\$E> **Now** six eighty so please thank you. <$2> Thank you.
<$E> sound of till <\$E> **Now** seventy eight so please. <$E> pause <\$E> Thank you very

[A]Conclusion

Spoken corpora were for a long time seen as appendages to much larger written collections of data. They have moved beyond this status and, aided by digital technology, more and more are emerging. It is an exciting point as we move towards the next generation of spoken corpora. Many challenges still remain, such as the drudge and expense of transcription, and challenges of audiovisual alignment with transcripts (see Adolphs and Carter, 2008). Adolphs and Knight (2010) note that as advances in technology allow us to develop new kinds of spoken corpora, such as audiovisual data-streams and much richer description of contextual variables, it will become increasingly important to agree on conventions for recording and representing these kinds of data, and the associated metadata. Adherence to agreed conventions, according to Adolphs and Knight, especially when developing new kinds of multi-modal and contextually-enhanced spoken corpora, will significantly extend the scope of spoken corpus linguistics into the future.

References

Adolphs, S. & Knight, D. (2010). Building a spoken corpus: what are the basics? In: A. O'Keeffe & M. J. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics.* (pp.38-52). London: Routledge.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999) *The Longman Grammar of Spoken and Written English.* Harlow, England: Pearson Education.

Carter, R. A. & McCarthy, M. J. (2006). *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Carter, R. & Adolphs, S. (2008). Linking the Verbal and Visual: New Directions for Corpus Linguistics. *Language and Computers* 64: 275–91.

Cheng, W. & Warren, M. (2000). The Hong Kong Corpus of Spoken English: Language Learning through Language Description. In: Burnard, L. & T. McEnery, (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective.* (pp. 133-144). Frankfurt am Main: Peter Lang.

Cheng, W., Greaves, C. & Warren, M. (2005). The creation of a prosodically transcribed intercultural corpus: The Hong Kong Corpus of Spoken English (prosodic). *International Computer Archive of Modern English (ICAME) Journal* 29: 5-26.

Cheng, W., Greaves, C., Sinclair, J. & Warren, M. (2009). Uncovering the Extent of the Phraseological Tendency: Towards a Systematic Analysis of Concgrams. *Applied Linguistics* 30 (2): 236–52.

Cook, G. (1990). Transcribing infinity: problems of context presentation. *Journal of Pragmatics* 14 (1): 1-24.

Crowdy, S. (1993).. Spoken corpus design. *Literary and Linguistic Computing* 8: 259-265.

Evison, J. (2010). What are the basics of analysing a corpus? In: O'Keeffe, A. & M. J. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics.* (pp. 122 – 135). London: Routledge.

Farr, F., Murphy, B. & O'Keeffe, A. (2002). The Limerick Corpus of Irish English: design, description & application. *Teanga* 21: 5-29.

Greaves, C. & Warren, M. (2010). What can a corpus tell us about multi-word units? In: A. O'Keeffe & M. J. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics.* London: Routledge, pp. 212 – 226.

Handford, M. (2010). *The Language of Business Meetings*. Cambridge: Cambridge University Press.

Harrington, J. (2010). *Phonetic Analysis of Speech Corpora*. Oxford: Wiley-Blackwell.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English based on the British National Corpus*. Harlow: Longman.

Martinez, R. & Schmitt, N. (under review). A phrasal expressions list.

McCarthy, M. J. (1998). *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

McCarthy, M. J. (2010). Spoken fluency revisited. *English Profile Journal* 1. Available at: http://journals.cambridge.org/action/displayJournal?jid=EPJ.

McCarthy, M. J. & Carter, R. A. (2002). *This that and the other*: Multi-word clusters in spoken English as visible patterns of interaction. *Teanga* 21: 30-52.

McCarthy, M. J. & Handford, M. (2004). 'Invisible to us': A preliminary corpus-based study of spoken business English. In U. Connor & T. Upton (Eds.), *Discourse in the Professions. Perspectives from Corpus Linguistics*. (pp. 167-201). Amsterdam, John Benjamins.

O'Keeffe, A., McCarthy, M. J. & Carter, R. A. (2007). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

O'Keeffe, A. & Walsh, S. (forthcoming). Applying corpus linguistics and conversation analysis in the investigation of small group teaching in higher education. *Corpus Linguistics and Linguistic Theory.*

Shin, D. & Nation, P. (2008). Beyond single words: the most frequent collocations in spoken English. *ELT Journal* 62 (4): 339-348.

Stenström, A-B., Andersen, G. & Hasund, I. K. (2002).*Trends in Teenage Talk. Corpus compilation, analysis and findings*. Amsterdam: John Benjamins

Tao, H. (2003). Turn Initiators in Spoken English: a Corpus-Based Approach to Interaction and Grammar. In: P. Leistyna & C. Meyer (Eds.), *Corpus Analysis: Language Structure and Language Use.* (pp. 187-207). Amsterdam: Rodopi.

Timmis, I. (2009). 'Tails' of linguistic survival. *Applied Linguistics* Advance Access. Available at: http://applij.oxfordjournals.org/cgi/search?fulltext=timmis+tails&x=14&y=4. Accessed 23.3.2010.

Thompson, P. (2010). Building a specialised audio-visual corpus. In: A. O'Keeffe & M. J. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics.* (pp. 93-103). London: Routledge.

Thornbury, S. (2010). What can a corpus tell us about discourse? In: A. O'Keeffe & M. J. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics.* (pp. 270 – 287). London: Routledge.

Warren, M. (2004). //↘ so what have YOU been WORKing on REcently//: Compiling a Specialized Corpus of Spoken Business English. In U. Connor & T. Upton (Eds.), *Discourse in the Professions: Perspectives from Corpus Linguistics*. (pp. 115–40). Amsterdam: John Benjamins.

Suggested readings

Adolphs, S. & Knight, D. (2010). Building a spoken corpus: what are the basics? In: A. O'Keeffe & M. J. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics.* (pp.38-52). London: Routledge.

Carter, R. & Adolphs, S. (2008). Linking the Verbal and Visual: New Directions for Corpus Linguistics. *Language and Computers* 64: 275–91.

Cook, G. (1990). Transcribing infinity: problems of context presentation. *Journal of Pragmatics* 14 (1): 1-24.

Crowdy, S. (1993).. Spoken corpus design. *Literary and Linguistic Computing* 8: 259-265.

Evison, J. (2010). What are the basics of analysing a corpus? In: O'Keeffe, A. & M. J. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics.* (pp. 122 – 135). London: Routledge.

McCarthy, M. J. (1998). *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

O'Keeffe, A., McCarthy, M. J. & Carter, R. A. (2007). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

To consult frequency data from the American National Corpus, see:
http://www.americannationalcorpus.org/frequency.html#