EL4252: Honours Year 2019/20

Session No. 2 : Data and transcription

1. Quantitative analysis *v.* qualitative analysis. The notion of 'triangulation' (metaphor from land survey – OED: 'The tracing and measurement of a series or network of triangles in order to survey and map out a territory or region, *spec.* by measuring the angles and one side of each triangle').

2. Natural (naturally occurring) data; corpus data; informant data. Researcher as interviewer, bystander, eavesdropper. The Observer's Paradox.

3. The web as a resource – eg MICASE: Michigan Corpus of Academic Spoken English[1] http://quod.lib.umich.edu/m/micase/
Other corpora of English (unfortunately, many require a licence fee) – non-definitive!
- Australian National Corpus https://www.ausnc.org.au/ ('The aim of the Australian National Corpus is to serve as a trusted service to collect, store, and provide access to a wide range of samples of Australian language for use in academic research')
- BNC (British National Corpus) http://www.natcorp.ox.ac.uk/
- Bank of English http://www.titania.bham.ac.uk/docs/svenguide.html
- Bergen Corpus of London Teenage Language (COLT) http://torvald.aksis.uib.no/colt/
- Brown, LOB (1961); Frown, FLOB (1991)
- Cambridge International Corpus (CIC) http://www.cambridge.org/sg/cambridgeenglish/about-cambridge-english/cambridge-english-corpus
- Corpus of Contemporary American English (COCA): http://corpus.byu.edu/coca/
- International Corpus of English (ICE) http://www.ucl.ac.uk/english-usage/ice/
- Santa Barbara Corpus of Spoken American English http://www.ldc.upenn.edu/Projects/SBCSAE/

4. The UK parliament site: http://www.parliament.uk/hansard/hansard2.cfm. Live video and a 28-day archive is available from http://www.parliamentlive.tv/. Some articles:
- Sandra Harris (2001), 'Being politically impolite: extending politeness theory to adversarial political discourse', *Discourse & Society* 12/4, 451–472
- Soledad Pérez de Ayala (2001), 'FTAs and Erskine May: Conflicting needs? – Politeness in question time', *Journal of Pragmatics* 33, 143–169.

   Also,
- Australian hansard: http://www.aph.gov.au/hansard/
- Canadian parliament: http://www.parl.gc.ca/common/index.asp?Language=E&Parl=37&ses=1
- New Zealand hansard: http://www.knowledge-basket.co.nz/gpprint/hansard.html
- Singapore hansard: https://www.parliament.gov.sg/parliamentary-business/official-reports-(parl-debates)

5. Wordsmith tools available in the Semiotics Lab (LRS 'Laboratory for Research in Semiotics', level 2, AS5). Freeware download: go to http://www.liv.ac.uk/~ms2928/wordsmith/

   NOTE: Access to the LRS is by your smartcard/matric card. Also installed in all the computers are various corpora. Go to the folder in the c:/ drive called 'Corpus Data'. You will find ICE-SIN, ICE-GB, LOB and other corpora. ICE-SIN is the Singapore component of the International Corpus of English; ICE-GB is the British component); LOB is the Lancaster-Oslo/Bergen Corpus (see http://helmer.aksis.uib.no/icame/lobman/lob-cont.html).

A. **Initial task**

B. **Second activity: shorter passages**

6. A manipulation of the **spelling** and **punctuation** (graphological features):
   (*a*) sometimes used to indicate pronunciation, tone, loudness (linguistic and paralinguistic features);
   (*b*) sometimes used to indicate attitudes;
   (*c*) sometimes used to indicate grammar or speech act.

7. Problem: (*a*), (*b*) and (*c*) can sometimes be confused; and the respelling, etc. is not always **consistent**. These might constitute **eye dialect** to the reader of the transcript, and therefore indicate the transcriber's **attitudinal judgement** of the interlocutor in question.

8. My recommendation: **conventional spelling** should almost always be used; if we need to indicate a divergent pronunciation, etc., use the IPA.

9. Conventional **punctuation** is problematic because these are not necessarily 'heard', and are used to indicate grammatical structure in writing.

---

10. There are **special features** of speech that are not captured in conventional writing. In particular, there are what are called **paralinguistic features**.

Here is a definition from Katie Wales's *A Dictionary of Stylistics* (2nd edn):

Definitions of **paralanguage** vary considerably, according to what is included or excluded. But generally it is recognised that COMMUNICATION in the spoken MEDIUM involves not only the UTTERANCES that realise language (VERBAL), but also other systems of SIGN also, that are NON-VERBAL (q.v.).

In this respect paralanguage is often regarded as a non-verbal, but vocal system, along with PROSODIC FEATURES (q.v.) such as pitch and loudness. And characteristic **paralinguistic features** would be noises that do not function as PHONEMES (i.e. in building words), but nonetheless do communicate a 'meaning' or ATTITUDE in speech: eg giggles, snorts, exclamations of disgust, disapproval, boredom, etc. But other definitions or discussions would include prosodic features also (see Elam 1980); and still others non-vocal signs like facial expressions and gestures, hence virtually synonymous with NON-VERBAL COMMUNICATION (q.v.) (see Lyons 1977).

Paralanguage significantly interacts with language in spoken discourse in ways that DISCOURSE ANALYSTS are keen to study. It is not easily represented in the GRAPHIC MEDIUM of novelistic DIALOGUE . Speakers rely on paralinguistic FEEDBACK from their ADDRESSEES; the audience watching and listening to a play can catch a whole range of emotional and attitudinal CONNOTATIONS from the vocalisations of the actors. They in turn can judge from the laughs, boos, hisses or coughs something of the audience's reactions to their own performances. [pages 282–283]

Three things isolated that need representing:
(a) **non-words** ('noises that do not function as phonemes');
(b) **prosodic features** (loudness, pace, pitch, intonation, stress, etc.); and
(c) **non-verbal features** ('body language').

Another feature not mentioned is the presence of **overlapping** or **interrupting** sequences, which will also need to be represented. There will also be **gaps** or **pauses** that will need to be taken into account.

11. Ultimately, we are free to choose any method for transcription, provided it is
(a) **consistent** (and therefore, not confusing or ambiguous);
(b) **economical** (calls for the least effort from the transcriber and reader); and
(c) **elegant** (does not look too off-putting).

From Dubois's (1991) list of transcription design principles:
Category definition: *define good categories*
Accessibility: *make the system accessible*
Robustness: *make representations robust* (i.e., use available characters, avoid invisible contrasts, avoid fragile contrasts)
Economy: *make representations economical*
Adaptability: *make the system adaptable* (integration of features for other purposes)

From Edwards (2008).

1. **Proximity of related events**: Events or types of information which are more closely related to each other are placed spatially nearer to each other than those which are less closely related. For example, prosodic information, such as prominent syllable stress, is often indicated by a mark (*eg* an apostrophe or an asterisk) placed immediately before the relevant syllable (cf. Svartvik and Quirk 1980; Gumperz and Berenz 1993).

2. **Visual separability of unlike events**: Events or types of information which are qualitatively different from each other (*eg* spoken words and researcher comments, codes, and categories) tend to be encoded in distinctly different ways. For example, codes may be enclosed in parentheses, or expressed as non-alphabetic characters (rather than alphabetic) or upper case letters (in contrast to lower case). This enables the reader to know what kind of information is about to be read before actually reading it, and thereby speeds reading and minimizes false attributions (*eg* perceiving a word as having been part of the speech stream, when it was really part of a metacomment or code).

3. **Time-space iconicity**: Temporally prior events are encountered earlier on the page (top to bottom or left to right) than temporally later events. This can include utterances, gestures, door slams, laughs, coughs, and so forth.

4. **Logical priority**: Logically prerequisite information for interpreting utterances tends to be encountered earlier on the page than the utterance(s) for which it is relevant. Information concerning the circumstances of data gathering and the relationships among the speakers tends to be given at the top of the transcript, whereas changes in circumstances or activities during the course of the interaction tend to precede the utterances they contextualize or potentially influence.

5. **Mnemonic marking**: Coded categories are encoded either in directly interpretable abbreviations or in symbolically iconic ways in order to expedite recovery of their meaning during rapid reading. An example of this is the use of a slash (/) for rising intonation and a backslash (\) for falling tone, rather than vice versa or instead of an arbitrary numerical code (*eg* '7'), as in the following example:

    (1) London-Lund Corpus, text 1.3 (Svartvik and Quirk 1980):

    13 212280 1 1 A 11 and at ^ h\ /ome#. /

    13 7212290 1 1 A 11 she's not a ^ b\ it the way she is at c/ollege# /

    Direct readability is also helped by using conventions already known from other written contexts. Du Bois (1991) notes that a number of transcription conventions derive from literary conventions found in novels and plays. Some examples are the use of three dots (…) for pauses, or a dash (−) for interrupted thoughts or utterances.

6. **Efficiency and compactness**: Coded distinctions should be marked with as few symbols as possible (*eg* non-redundantly, using short abbreviations), so long as meaning is easily recoverable (*ie* encoded mnemonically). This serves to minimize nonessential and distracting clutter in the transcript. For example, the use of a slash (/) for rising tone is more compact and efficiently read than

would be the use of the full word, 'rising'. The encoding of spoken words and prosodic information on the same line instead of on separate lines is also a type of compactness.

12. We also want to distinguish between a **narrow** transcription (with many details) and a **broad** transcription (with few details).

13. We can examine transcripts based on their treatment of the following:
   (*a*) overlaps,
   (*b*) gaps/pauses,
   (*c*) undecipherable sounds or guesses,
   (*d*) tone units [see below],
   (*e*) 'lazy' or 'inaccurate' pronunciations,
   (*f*) tone/pitch,
   (*g*) stress,
   (*h*) intensity (loudness),
   (*i*) voice quality,

   (*j*) non-words (*er*/*uh* or Ə or ɜː or [vocalised pause]), and
   (*k*) non-verbal features.

> **tone group/unit** A term used by some INTONATION analysts, particularly those working within the British tradition, to refer to a distinctive sequence of PITCHES or TONES, in an utterance. The essential feature of a tone unit is the NUCLEAR tone, the most PROMINENT tone in the sequence; and this may be accompanied, depending on the length of the utterance, by other components, such as the **head** (i.e. the sequence of SYLLABLES between the first STRESSED syllable and the nuclear tone), **pre-head** (i.e. unstressed syllables at the very beginning of the tone unit) and **tail** (i.e. the syllables following the nuclear tone). This terminology can be illustrated in the sentence *the | man `bought a | nèw | `clock |*, where the sequence of pre-head/head/nucleus/tail/tail is marked by vertical lines. A tone unit usually corresponds to a CLAUSE or SENTENCE, but may be used on any grammatical unit, *eg* in an extremely irritated version of the above sentence, there might be several tone units, as in *the mán | bóught | a néw | clòck |*. See Crystal 1969: Ch. 4; Gimson 1989: Ch. 10. [David Crystal, *A Dictionary of Linguistics and Phonetics*, p. 353.]

14. There is also the **ethical problem** of recording others' speech, especially if this is done **surreptitiously**. If recording is done openly, there is the problem of **authenticity**. The desire is for **naturally occurring** data, but many claim that the language which linguists would most like to record is the language which is most susceptible to *contamination by observation* — the **observer's paradox**. The data is said to be **tape-affected** and **researcher-affected**. (See issues raised in the Hammersley article.)

15. An examination of some transcriptions (see Appendix).

16. One possible way of doing it:

1. Prosodic Features
(a) The <u>colon</u> indicates a lengthened vowel; additional colons can be used to indicate increased length – *eg* 'so:: are you re:ally doing it'
(b) The <u>question mark</u> indicates 'questioning' intonation – it does not necessarily reflect the grammatical structure of the utterance – *eg*, 'then?'
(c) The <u>full stop</u> indicates 'sentence final' intonation — again it does not necessarily reflect grammatical completeness – *eg*, 'so. that's it. for now anyway.'
(d) <u>Underlining</u> indicates stress– *eg*, 'I never said that.'
(e) <u>Upper-case letters</u> indicate loudness – *eg*, 'So I told her GO AWAY'
(f) The <u>exclamation mark</u> indicates a high-fall intonation – *eg*, 'It was so silly!'

2. Timing signs
(a) The <u>hyphen</u> can be used to indicate a pause of about ½ second; therefore, two hyphens for 1 second, three hyphens for 1½ secs., etc. (Separate the hyphens with spaces on either side so that it will be obvious these are timing symbols.)
(b) <u>Equal signs</u> between words in a line indicate that a person spoke very fast – *eg*, 'I=don't=know'
(c) <u>Paired equal signs</u> at the end of one line and at the beginning of the next indicate that there is no pause in the flow of speech – *eg*,
   John: When are you coming.=
   Mary: =Don't know lah.
(d) <u>Paired single and double asterisks</u> between two lines of text indicate simultaneous speech — *eg*,
   John: When are you *coming.*
   Mary: *Don't know* lah I really cannot say now 'cos I'm so **busy now**
   John: **Just tell** me roughly I need to know - so when.
   John says 'coming' when Mary says 'Don't know'; Mary says 'busy now' when John says 'Just tell'.

3. Signs indicating inaudibility, etc.
(a) <u>Words enclosed in single parentheses</u> indicate that what is said is not clear and the words are the transcriber's guesses – *eg*, 'It's really quite (late)'
(b) <u>Empty parentheses</u> indicate that something was said, but was inaudible or uninterpretable – *eg*, 'It's really quite (   )'; you can also indicate the number of syllables spoken thus: 'It's really quite (1 syllable)'

4. Extra-textual information
(a) <u>Square brackets</u> enclose transcriber's remarks – *eg*, 'When can we makan ['eat', Malay]'
(b) <u>Square brackets</u> also enclose paralinguistic information not accounted for by the prosodic diacritics – *eg*, 'No! [laugh] you don't mean it!'

17. Important note: don't use abbreviations, symbols, pictographs, etc. — in other words, if someone says 'See you a half past seven tomorrow', don't transcribe 'See you at 7.30 pm tomorrow'; or if someone says, 'I bought that house for one and a half million', don't write 'I bought that house for S$1,500,000.'

18. Related issue: where can I record data?
- at home (not necessarily your own)
- in a café or restaurant or canteen
- in school or university
- film (online, DVD, VCD), television and radio (*scripted speech may not be appropriate to your study – remember that if you analysed language used in a sit-com or a comedy as naturalistic language, your analysis will be highly problematic!*)

C. **Short Exercise**

D. **Transcription activity**

Also see:
Appendix 1
Appendix 2
Appendix 3
Appendix 4
Appendix 5

READINGS

Bucholtz, Mary (2000), 'The politics of transcription', *Journal of Pragmatics* 32, 1439–1465. (Available online through library website.)

*Deborah Cameron (2001), *Working with Spoken Discourse* (London: Sage), Chh 2 and 3. [In LumiNUS]

*Edwards, Jane A (2003), 'The Transcription of Discourse', in D Schiffrin, D Tannen & H E Hamilton, *The Handbook of Discourse Analysis* (Oxford: Blackwell, 2003), Ch 17. [In LumiNUS]

Hammersley, Martyn (2003), ' "Analytics" are no substitute for methodology: a response to Speer and Hutchby', *Sociology* 37(2): 309–351.

Langford, David (1994), *Analysing Talk: Investigating Verbal Interaction in English* (London: Macmillan, 1994), Ch. 3.

Schiffrin, Deborah (1994), *Approaches to Discourse* (Oxford: Blackwell, 1994), Appendixes 1 and 2 (pp. 420-438)

Stubbs, Michael (1983), *Discourse Analysis* (Oxford: Blackwell, 1983), Ch. 11.

Wray, Alison *et al.* (1998), *Project in Linguistics: A Practical Guide to Researching Language* (London: Arnold, 1998), Chs. 12 and 18.

Back to EL4252 contents page