Critical issues in spoken corpus development: the Spoken BNC2014 transcription scheme and speaker identification

Robbie Love

CASS, Lancaster University

BRITISH NATIONAL CORPUS 2014

Lancaster University

@lovermob

Today's talk

CASS
Corpus Approaches to Social Science

- The BNC
- The Spoken BNC2014
- Progress so far – pilot study + current work
    - (1) Transcription scheme development
    - (2) Speaker identification
- Conclusions
- Next steps

Some history: the BNC (1994)

CASS
Corpus Approaches to Social Science

"a corpus of 100 million words of written texts and spoken transcriptions of modern British English, to be stored on the computer in machine-readable form"

Leech (1993: 9)

- British publishers: Oxford University Press, Longman, Chambers
- Plus Oxford University and Lancaster University

Some history: the BNC (1994)

CASS
Corpus Approaches to Social Science

- BNC used to produce over 200 journal articles (over 100 published after 2009)

- Open-access, hosted online by various institutions:
  - Brigham Young University (BNC-BYU)
  - University of Zurich (BNCweb World Edition)
  - Lancaster University (BNCweb)

Some history: the BNC (1994)

CASS
Corpus Approaches to Social Science

- Spoken component = 10 million words
- **Demographic** (c. 40%) and context-governed data (c. 60%) (see Aston and Burnard 1997)

"an immense collection of conversational data, systematically sampled across the whole population of the UK…a comprehensive and carefully sampled record of how the language is used in living speech"

(Leech 1993: 14)

Some history: the BNC (1994)

CASS
Corpus Approaches to Social Science

- It's getting old
- Can no longer be used as a proxy for present day British English

- Nothing since the Spoken BNC (1994):
  - large size
  - general coverage of spoken British English
  - (low or no cost) public access
  - transcribed

The BNC2014

CASS
Corpus Approaches to Social Science

The Spoken BNC2014

- CASS and Cambridge University Press
- 10 million words spontaneous conversation (*demographic* data)

- First of its kind since the original Spoken BNC (1994)

People

- Claire Dembry
- Laura Grimes
- Samantha Owen
- 13 transcribers

- Tony McEnery
- Andrew Hardie
- Vaclav Brezina
- Robbie Love

The Spoken BNC2014

- See Dembry and Love (2015) for overview of methodology

Some highlights:

- Members of the public commissioned as freelancers to make unsupervised recordings
- Smartphones (vs. analogue tape recorders)
- Non-surreptitious!
- 13 freelance transcribers
- Only demographic (for now)

The Spoken BNC2014

Both parties

- Fund project equally
- Encourage participation – media campaigns
- Disseminate information

CUP

- Corresponds with contributors
- Collects recordings
- Transcribes data

CASS

- Carries out methodological investigations
- Converts transcripts to XML, encoding
- Annotates corpus
- Initial analysis
- Prepares for public release/hosts finished corpus

The Spoken BNC2014

CASS
Corpus Approaches to Social Science

Progress report

- **4 million** words transcribed
- Average 140,000 words per week

- 373 recordings submitted
- 300 hours submitted
- 367 speakers so far and counting

The Lancaster pilot study (Love 2014)

- Investigation of several methodological issues in the compilation of spoken corpora – informing practice in the Spoken BNC2014
  - Design and metadata
  - Recording
  - **Transcription**
    - **developing the scheme**
    - **Speaker identification**

- Simultaneous to the Cambridge pilot study (Dembry)

The Lancaster pilot study (Love 2014)

CASS
Corpus Approaches to Social Science

The Spoken BNC2014 pilot corpus

- 5.5 hours of audio data
- Replicated the style of recordings in the Spoken BNC2014
- 14 recordings, 32 speakers, 47,000 words, 6,552 turns
- Transcribed by two full-time, professional transcribers at CASS

Transcription scheme development

- Crowdy (1994) – *original BNC scheme*
- Gablasova et al. (under review) – *Trinity Lancaster Corpus*
- Atkins et al. (1992) – *corpus design criteria*
- Hasund (1998) – *anonymization guidelines*
- Consultation with CASS transcribers
- Claire Dembry's work at Cambridge (2012-)
- Discussion between CASS and CUP
- Hardie (2014) - XML

Why not simply reuse the original?

CASS
Corpus Approaches to Social Science

Crowdy (1994) “Spoken Corpus Transcription”

Generally, it’s pretty good, but:

- 16 features identified in the **1,900** word scheme – very few examples
- Not enough clarity in some areas, leading to ambiguity
- Compatibility with CASS XML standards for automatic conversion

Why not simply reuse the original?

CASS
Corpus Approaches to Social Science

EXAMPLE #1

- Question marks to indicate questioning utterances

<1> It's a funny old day isn't it.

<2> Mm it's not cold is it?

Crowdy (1994: 28)

Why not simply reuse the original?

CASS
Corpus Approaches to Social Science

EXAMPLE #2

- using full stops and commas to “approximate to use in written text”, but also indicating pauses with ellipses

<2> I think it’s always, deceptive on days like this because its, overcast and [er]

[…]

<2> But, but er, he’s…just broken away from his girlfiend and [<unclear>]

<1> [Oh has] he, oh. Well he seemed happy enough when he called.

Crowdy (1994: 28)

Trinity Lancaster corpus scheme

- Transcription scheme for the Trinity Lancaster Spoken Learner Corpus (Gablasova et al. under review)
- One-to-one, examiner-student conversations
- But major advantage of modernity, CASS-compatibility and success in Trinity Lancaster project

Trinity Lancaster corpus scheme

- Started with Gablasova et al.'s scheme
- Adapted according to
  - Crowdy (1994), and
  - Atkins et al. (1992: 11-12), who provide a nice and still useful set of recommended considerations

Revising the Trinity Lancaster scheme

CASS
Corpus Approaches to Social Science

Encoding of speaker IDs

- Speakers assigned unique numeric label (Crowdy 1994)

<1> It's a funny old day isn't it.

<2> Mm it's not cold is it?

Crowdy (1994: 28)

Revising the Trinity Lancaster scheme

CASS
Corpus Approaches to Social Science

Anonymization

- Omit "any reference that would allow an individual to identified" (Crowdy 1994)
- NOT automatically (Hasund 1998)
- Hasund: Bank of English includes gender in anonymization tag

e.g. *I bumped into <name female> yesterday*

(+ male, neutral)

Revising the Trinity Lancaster scheme

CASS
Corpus Approaches to Social Science

Overlaps

- Crowdy's (1994) rather complicated system:

<1> So she was virtually a [a house prisoner]

<2> [house {bound}]

<3> {prisoner}

- Not in Trinity Lancaster scheme
- Decision to retain omission

Revising the Trinity Lancaster scheme

CASS
Corpus Approaches to Social Science

Punctuation

- Too much room for interpretation in Crowdy (1994)
- Like Trinity Lancaster corpus, all syntactic punctuation stripped, apart from question marks for *fully formed* interrogative structures
  - *yes/no questions*
  - *wh-questions*
  - *tag questions*

Revising the Trinity Lancaster scheme

CASS
Corpus Approaches to Social Science

Quotative speech

- Not in Atkins et al. (1992), Crowdy (1994) or Trinity Lancaster scheme
- Proposal:

<1> he said <quot> I'll see you later </quot>

- Could this be added to scheme with minimal time addition?

Pilot with CASS transcribers

- Tested in the transcription of the Spoken BNC2014 pilot corpus
- Consultation with the CASS transcribers, who were also the transcribers on the Trinity Lancaster project
- Further changes made in reflection

Pilot with CASS transcribers

CASS
Corpus Approaches to Social Science

Anonymization

- Of the 380 <*name*> tags, only 1.8% not coded for gender

Pilot with CASS transcribers

CASS
Corpus Approaches to Social Science

Question marks

- Crowdy (1994) criticised for being too loose with this
- However CASS transcribers wanted more than fully formed interrogatives
- Trusted to use intuitive criteria instead, e.g.

<3> ah is it lovely and warm there Dylan? getting dried off?

<?> how many years have we lived here? two and a half years?

Pilot with CASS transcribers

CASS
Corpus Approaches to Social Science

Quotative content

- CASS transcribers reported no problems
- But, e.g., only 35/75 instances of *said + direct reported speech* actually tagged
- Therefore removed from scheme

Review with Cambridge

- Resulting scheme sent to CUP to ‘merge’ with scheme used so far by their team
- Features added that were not considered by previous Lancaster investigations but deemed worthy of inclusion
  - *filled pauses*
  - *non-English speech*
  - *pauses*
  - *events*

Resulting scheme

CASS
Corpus Approaches to Social Science

- From 1,900 words to **5,000 words!**
- Lots of examples
- (Hopefully) minimal room for ambiguity = maximal room for inter-rater consistency

The bird's eye view

SPOKEN BNC (1994)	SPOKEN BNC2014
Speaker turns	Speaker IDs
Overlapping speech	Overlaps
Use of punctuation, and 'sentence' boundaries	Punctuation – question marks
	Utterances
	Unfinished words (false starts)
Pauses	Pauses and events
Vocalised pauses	Pauses and events
Accent, dialect, and representation of nonstandard forms	Nonstandard words or sounds
	Nonstandard contractions or shortenings
	Native speaker accent/dialect
Paralinguistic features	Pauses and events
Non-verbal sounds	Non-linguistic vocalisations
Contextual comments	Pauses and events
Unclear or inaudible text	Unintelligible speech/guesses
Unfamiliar words	Unintelligible speech/guesses
Spelt-out words	Acronyms/spelling/capitalisation
Acronyms and abbreviations	Acronyms/spelling/capitalisation
Telephone conversations	Pauses and events
Codes used to preserve anonymity	Anonymization
Text read out	Pauses and events

EXTRA SPOKEN BNC2014
General guidelines
Document format
Line height and spacing
Header information
Tag format
Non-English speech
Numbers

The bird's eye view

CASS
Corpus Approaches to Social Science

- Delicate balance sought between
  - backwards compatibility, and
  - optimal practice
- Similar enough to compare with original
- Different enough to be better

eXtensible Markup Language (XML)

CASS
Corpus Approaches to Social Science

- Makes possible for automated mapping to standard XML, with minimal manual editing
- Original Spoken BNC was not initially in XML, but later converted, therefore comparable
- But even in XML it adheres to the highly complex Text Encoding Initiative (TEI)

- So we're using Hardie's (2014) "Modest XML for Corpora"

"any linguist from the level of a bright undergraduate upwards should be able to understand it"

(p. 79)

A fresh problem

- Quality control traditionally focusses on accuracy of *transcription*
- Spoken BNC2014 is no exception – audio-checking and proofreading procedures in place at Cambridge

- However…a fresh problem arose in the Lancaster pilot study

Speaker identification
= who said that?

CASS
Corpus Approaches to Social Science

- has no bearing on the accuracy of the transcription of linguistic content itself (i.e. what was said), but refers to the identification of the speaker that produced the transcribed turn (i.e. who said it)

Speaker identification

- There are two unavoidable deficiencies in the transcription of audio recordings: transcribers' lack of familiarity with

  (i) the speakers and

  (ii) the context in which the conversations occurred

Major assumption

Speaker identification not an issue

- when there are only two speakers; or,
- when the speakers have highly contrasting voice qualities; or,
- when the transcriber knows the speakers in the recording, and can recognise their voices.

Major assumption

Speaker identification is likely an issue

- **when there are several speakers**, and/or
- when the differences in voice quality between two or more speakers are not be sufficient to tell them apart

Importance of speaker identification

CASS
Corpus Approaches to Social Science

- Speaker ID codes link to demographic metadata
- Corpus-based sociolinguistics is already controversial – aggregate data

"random (and therefore sociolinguistically irrelevant) speaker groupings can often yield statistically significant results"

Brezina & Meyerhoff (2014)

Pilot study (Love 2014)

- Pilot #1 = the Spoken BNC2014 pilot corpus
  - *certainty*

- Pilot #2 = legitimate Spoken BNC2014 recording
  - *certainty*
  - *inter-rater agreement* with original transcript

Pilot #1

Pilot #1 = *certainty* in the Spoken BNC2014 pilot corpus

- 5.5 hours of audio data
- Replicated the style of recordings in the Spoken BNC2014
- 14 recordings, 32 speakers, 47,000 words, 6,552 turns
- Transcribed by two full-time, professional transcribers at CASS

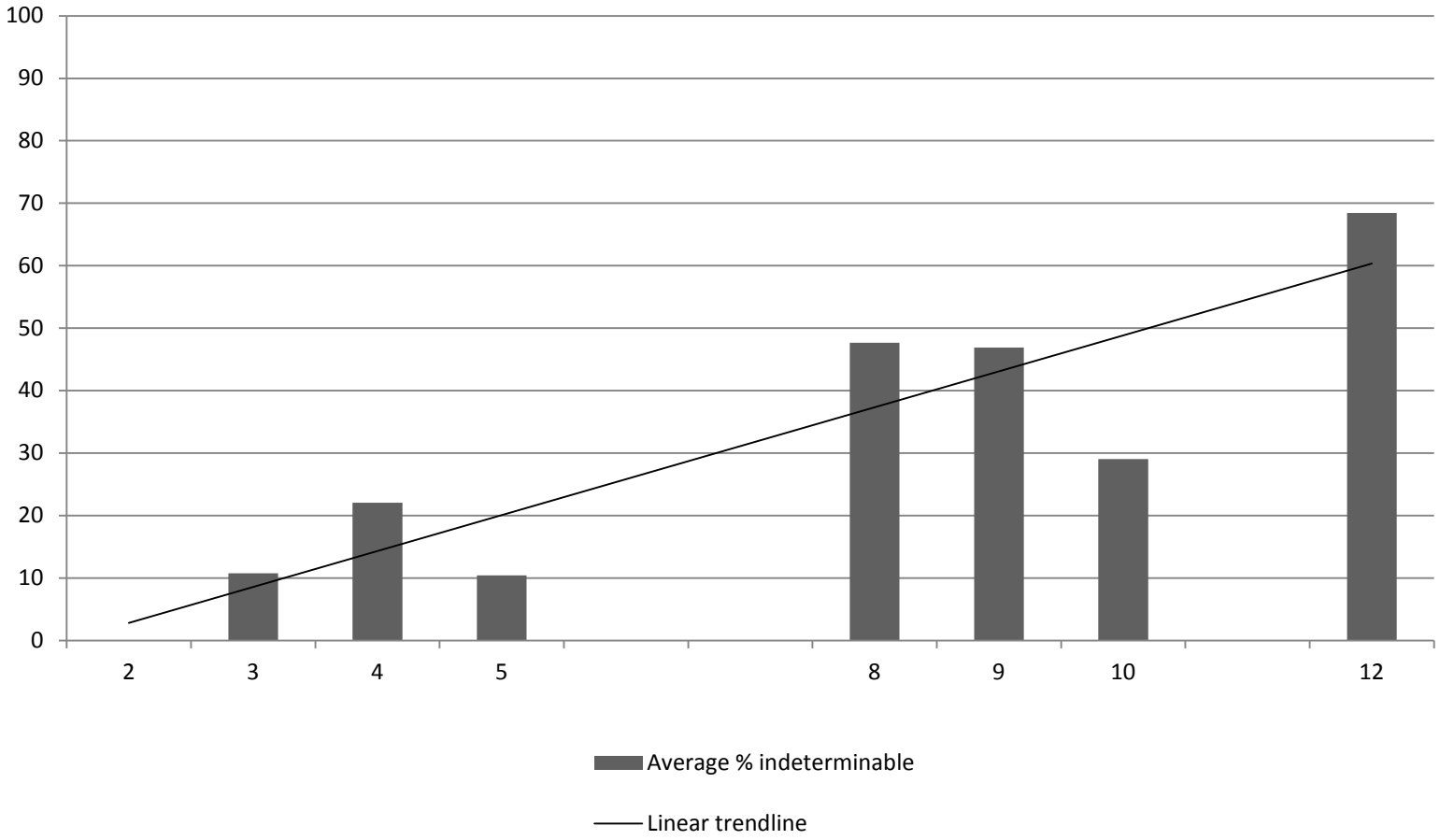
Certainty (the pilot corpus)

Speaker identification action	Example speaker ID code	% of turns in pilot study recordings
Mark turn with speaker ID code	<022>	68.31
Mark turn with 'best guess' speaker ID code	<022?>	6.26
Mark turn as indeterminable	<?>	25.43

- Certainty a majority, but 25% indeterminable
- No. of speakers?

Certainty (the pilot corpus)

CASS
Corpus Approaches to Social Science



Pilot #2

Pilot #2 = *certainty* and *inter-rater agreement* in a legitimate Spoken BNC2014 recording

- An example of the most difficult circumstance in the Spoken BNC2014 itself
- 9 speakers
- 1,080 turns
- 9,871 words
- Spoken BNC2014 transcript compared with two CASS transcribers' versions
- One used 'best guess', one didn't

Certainty (legitimate Spoken BNC2014 recording)

Speaker identification action	Example speaker ID code	% of turns in Spoken BNC2014 transcript	% turns in CASS transcriber version #1	% turns in CASS transcriber version #2
Mark turn with speaker ID code	<022>	94.35	18.06	70.09
Mark turn with 'best guess' speaker ID code	<022?>	0.28	42.78	0.00
Mark turn as indeterminable	<?>	5.37	35.56	23.70
Not coded	*N/A*	*0.00*	*3.61*	*6.20*

Certainty (legitimate Spoken BNC2014 recording)

- Was the original transcriber really as certain as the transcript implies?
- Speaker identity in this recording appears to be far from clear

- What about inter-rater agreement?
- 1,019 turns in original transcript had speaker ID codes

Inter-rater agreement (legitimate Spoken BNC2014 recording)

Type of match		% turns in CASS transcriber version #1	% turns in CASS transcriber version #2
Match	Exact	15.09	38.55
	Best guess	17.54	N/A
Non-match	Wrong code	32.16	35.77
	Indeterminable	35.21	25.67

Inter-rater agreement (legitimate Spoken BNC2014 recording)

- When code given (i.e. ignoring indeterminable codes) chance of matching only just over half
- However, 99.4% of wrong codes at least got the gender right
- So, in BNC2014 transcription scheme, *indeterminable* replaced with minimum gender code (i.e. <M> or <F>)

Current work

- Speaker identification could be an problem worth paying attention to
- Further investigation needed, within a reasonable limit
- ‘Speaker-heavy’ recordings = 20% of Spoken BNC2014 so far
- *ASSUMPTION: this is not a problem for 2-, 3-, 4-speaker recordings – to be checked!*

Current work

- Assessing the actual Spoken BNC2014 transcribers (rather than CASS transcribers)

- Replicating pilot work on Spoken BNC2014 transcribers, plus:
- Is there a ‘gold standard’?
- Can one be manufactured illegitimately?

Current work

- Investigation #1 = legitimate Spoken BNC2014 recording
  - *certainty*
  - *inter-rater agreement* with original transcript

- Investigation #2 = fake gold standard recording
  - *accuracy*

Investigation #1

Investigation #1 = *certainty* and *inter-rater agreement* in a legitimate Spoken BNC2014 recording

- 6 speakers, 32 minutes, 587 turns, 6,862 words
- Original transcript + 6 'test' transcripts
- Average proportion of definite ID codes versus indefinite ID codes
- Agreement on coding of specific ID codes between transcripts

Investigation #2

Investigation #2 = *accuracy* in a fake gold standard recording

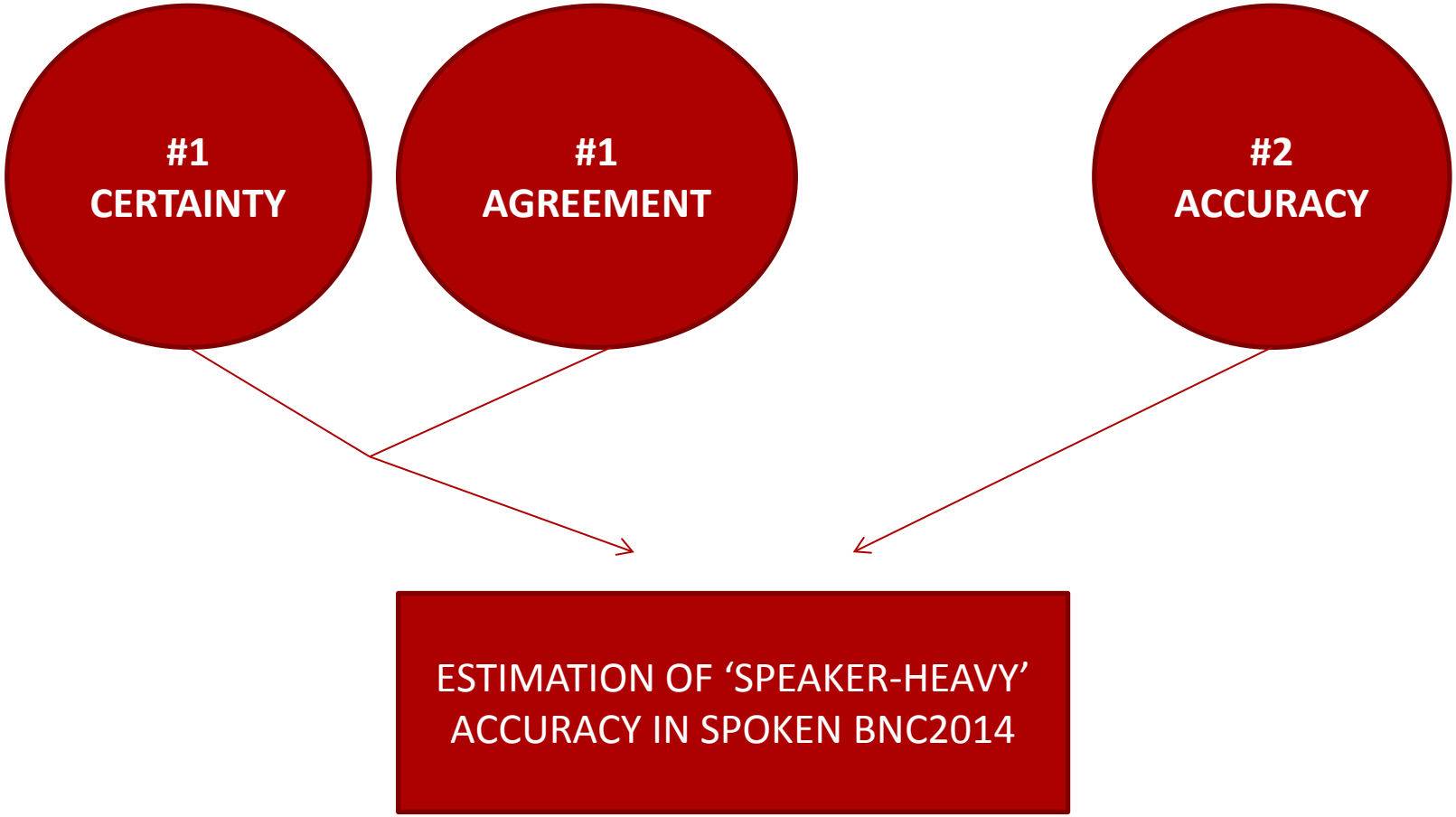
- 8 speakers, 25 minutes, 775 turns, 4,886 words
- My transcript ('gold standard') + 8 'test' transcripts
- Same as *inter-rater agreement* test, but we can call this *accuracy*

Preparing the data

- Both investigations asked the Spoken BNC2014 transcribers to transcribe the recordings from scratch
- #1 was done ‘blind’; #2 was done explicitly
- In both cases, transcripts had to be manually aligned

CASS
Corpus Approaches to Social Science

Current work: overview



Investigation #1 – findings so far

CASS
Corpus Approaches to Social Science

- Total 587 turns to compare

Speaker identification action	Example speaker ID code	% of turns in Spoken BNC2014 transcript	% turns in test transcripts
Mark turn with speaker ID code	<022>	98.30	80.99
Mark turn with 'best guess' speaker ID code	<022?>	1.53	0.56
Mark turn with gender	<F>	0.17	1.17

- Not yet aligned – so some transcripts have more turns than original

Investigation #2 – findings so far

CASS
Corpus Approaches to Social Science

- 775 turns considered across all 8 speakers

Transcriber	% accurate speaker ID
T01	38.87
T02	65.56
T03	51.24
T04	34.88
T05	76.99
T06	31.95
T07	70.45
T08	58.62
AVERAGE	**53.57**

Investigation #2 – findings so far

CASS
Corpus Approaches to Social Science

- 775 turns considered across all 8 speakers

Speaker	No. turns considered per speaker	% accurate speaker ID (all transcribers)
S01	122	72.44
S02	59	33.90
S03	62	**21.57**
S04	107	57.13
S05	115	55.98
S06	110	**83.52**
S07	79	33.54
S08	121	70.45
	775 (total)	**53.57 (average)**

Speaker identification – conclusions so far

CASS
Corpus Approaches to Social Science

- Variation between transcribers in certainty over same recording (Spoken BNC2014 transcript)

- Accuracy in gold standard only just over 50%, proving difficulty of this

Next steps

- Transcription development
  - Transcribe the rest and monitor audio-checking/proofreading procedure
- Speaker identification
  - Check that ‘speaker-light’ recordings (c. 80% of corpus) are not affected by this problem
  - Put appropriate warning label on finished corpus, ability to exclude ‘high risk’ recordings
  - Argue that this is worth paying attention to
  - Historical assessment of implications for previous spoken corpora/research

Participate!

- Data collection is ongoing

corpus@cambridge.org

References

- Aston, G. & Burnard, L. 1997. The BNC handbook. Exploring the BNC with SARA. Edinburgh: Edinburgh University Press.
- Atkins, A., Clear, J., & Ostler, N. (1992). Corpus Design Criteria. *Literary and Linguistic Computing, 7*(1), 1-16.
- Brezina, V., & Meyerhoff, M. (2014). Significant or random? A critical review of sociolinguistic generalisations based on large corpora. International Journal of Corpus Linguistics, 19(1), 1-28. doi:10.1075/ijcl.19.1.01bre
- Crowdy, S. (1994). Spoken Corpus Transcription. Literary and Linguistic Computing, 9(1), 25-28.
- Gablasova, D., Brezina, V., McEnery, T. & Boyd, E. (under review) Epistemic stance in spoken L2 English: The effect of task type and speaker style, submitted to Applied Linguistics.
- Hardie, A. (2014). Modest XML for corpora: not a standard, but a suggestion. ICAME Journal, 38, 73-103.
- Leech, G. (1993). 100 million words of English. English Today, 9-15. doi:10.1017/S0266078400006854