Aston University
BIRMINGHAM UK

Deriving a new COLT from the Spoken BNC2014

The case of teenage swearing

Robbie Love
Aston University
@lovermob

Anna-Brita Stenström
University of Bergen

COLT

The Bergen Corpus of London Teenage Language (COLT)

- Stenström et al. (2002)
- Recorded in 1993

Size

- Approx. 500,000 words
- 100 audio tapes
- 50 hours of recorded conversations

Participants

- 31 teenagers, aged 13-17, recruited through their schools
- Went on to record 50+ other speakers

Settings

- Mainly in and around school, but also private and public social spaces

Two versions of COLT

Around the same time, the British National Corpus was under compilation

In exchange for including COLT in the BNC, **the BNC team transcribed recordings for Bergen**

- These transcripts are what feature in the BNC

Bergen then checked and corrected the BNC transcripts, **inserting lots of material previously not transcribed** (c. 20%), producing a more accurate version

- These transcripts are what are available via CLARIN

More on this later!

Spoken BNC2014

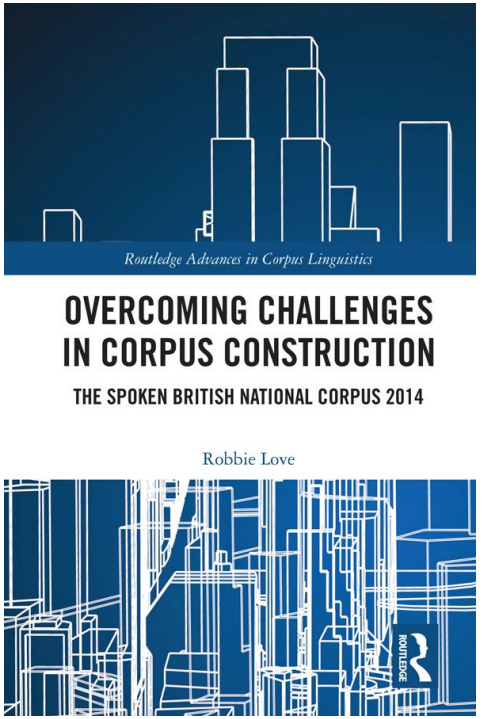
Spoken British National Corpora

Transcriptions of recorded conversations

1990s and 2010s

- Spoken BNC1994 (BNC Consortium, 2007)
  - c. 5 million words casual conversation
  - including COLT

- Spoken BNC2014 (Love et al., 2017)
  - c. 11 million words casual conversation



Deriving a ‘COLT 2’ from the Spoken BNC2014

Anna-Brita was interested in revisiting teenage swearing to see what might have changed since the 1990s

Asked about isolating teenage speakers in the Spoken BNC2014 for comparison to COLT data

Can a new equivalent of COLT be derived from the Spoken BNC2014?

Deriving a ‘COLT 2’ from the Spoken BNC2014

The Spoken BNC2014 was compiled in a very different context to COLT

- **COLT**
  - A separate study, explicitly targeting London teenage language
  - Recruits wore hidden tape recorders under their clothes
  - Incidental conversations
  - Torgersen et al. (2011): “COLT contains the speech of the friends, family and teachers of the recruits, with no consistent encoding of their age, sex, ethnicity and residence.”

- **Spoken BNC2014**
  - General corpus, trying to gather a broad sample of speakers
  - All speakers aware of recording; gave informed consent
  - Recording ‘sessions’ – conversation as an ‘event’
  - Sociolinguistic metadata for all speakers

Deriving a ‘COLT 2’ from the Spoken BNC2014

Aim

- Find texts in the Spoken BNC2014 to form a sub-corpus that is comparable to COLT

Questions

- What are the design criteria of the original COLT that would need to be present in a ‘new COLT’ in order to achieve comparability?
- Are there enough texts in the Spoken BNC2014 that meet these criteria so as to build a comparable sub-corpus?

Deriving a ‘COLT 2’ from the Spoken BNC2014

Are there enough teenage speakers from London in the Spoken BNC2014?

- Teenage speakers: 54 speakers aged 13-19
- But only 7 of them born in London

Expanded to more generous “south east England” category

  - This left 26 teenage speakers from “south east England”
  - BUT…some of them were talking to adults! We needed teenager-teenager conversations
  - **15 of these speakers** who featured in a total of **35 ‘teenager-only’ conversations** totalling **25 hours of recordings**
  - This allowed for the compilation of a sub-corpus by identification of the text IDs

Deriving a “COLT 2” from the Spoken BNC2014

Which platform/tool would be most appropriate for analysis?

- Simply upload the original COLT and the new COLT 2 to a tool for analysis?
- But isn’t COLT part of the original BNC? So why not access the BNC-COLT files?

Which version of the original COLT to compare against?

- **Original** version – part of the BNC
- **Edited** version – further checked and corrected by Bergen team, released as separate corpus:
  - “As a result, we have not only ended up with a transcription that is more faithful to the tape-recordings but also with a larger corpus; the number of words has increased by at least 15 per cent.”

(http://korpus.uib.no/icame/colt/COLTinfo.html)

An aside: finding COLT in the BNC1994?

We decided to try to isolate the original files in the BNC:

“The complete corpus, half a million words, has been orthographically transcribed and word-class tagged, and is a constituent of the British National Corpus.” (http://korpus.uib.no/icame/colt/)

- Problem: COLT files and BNC files use different naming conventions
- There does not seem to be a record of the correspondence between the conventions
  - Which BNC1994 files came from COLT?

An aside: finding COLT in the BNC1994?

Manual search of strings from each edited COLT file in the BNC1994

- e.g. COLT file **B132401**:

    <u who=1-1 id=33> What day is it, Thursday?

Search in BNC1994 for “What day is it”

Found in BNC file **KPT**:

    PS57T: What day is it , Thursday ?

With thanks to Catherine Wu (Aston University)

An aside: finding COLT in the BNC1994?

This resulted in the identification of 31 BNC1994 files that contained COLT material

However, unlike the BNC2014, a ‘text’ in the BNC1994 does not equate to a single ‘conversation’…

- e.g. BNC text KBD

> “24 conversations recorded by `Barry' (PS03W) between 1 and 6 February 1992 with 10 interlocutors, totalling 9021 s-units, 58087 words, and 5 hours 12 minutes 10 seconds of recordings.”

Some of this is COLT, but it is mixed in with non-COLT material

So, it has not (yet) been possible to properly isolate COLT within the BNC1994

Procedure

COLT 1

- For now, the edited COLT files were downloaded as a separate corpus from CLARIN: https://clarino.uib.no/korpuskel/clarino-metadata?identifier=colt
- Uploaded to Sketch Engine

COLT 2

- Spoken BNC2014 files downloaded from Lancaster University: http://corpora.lancs.ac.uk/bnc2014/signup.php
- Qualifying texts isolated and uploaded to Sketch Engine

Corpus data

- **COLT 1**
  - 83 speakers
  - 377 texts
  - c. 600,000 tokens
- **COLT 2**
  - 15 speakers
  - 35 texts
  - c. 300,000 tokens

Case study: swearing

Building on work of Stenström et al. (2002) (*Trends in teenage talk*) and Stenström (2006) (*Taboo words in teenage talk*)

1. Comparison of swear word frequencies:
- Between COLT 1 and COLT 2
- Between COLT 1 and Spoken BNC1994
- Between COLT 2 and Spoken BNC2014

2. All instances of FUCK retrieved and manually coded:
- Syntax
- Function (McEnery et al., 2000)

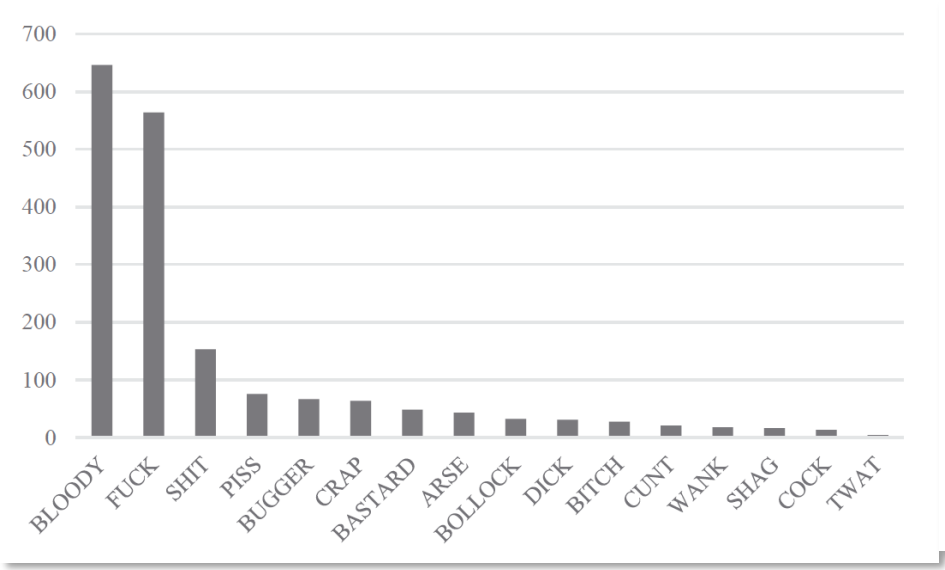
  - COLT 1 = 727 hits (652 per million words)
  - COLT 2 = 171 hits (547 per million words)

Categories of insult (McEnery, 2005: 27)

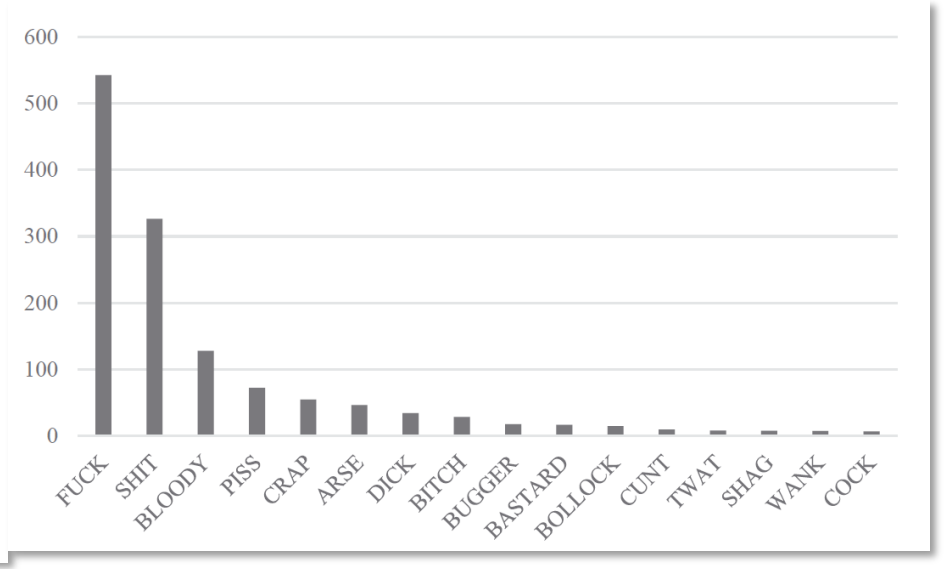
Letter	Code	Description
A	PredNeg	Predicative negative adjective: ‘the film is shit’
B	AdvB	Adverbial booster: ‘Fucking marvellous’ ‘Fucking awful’
C	Curse	Cursing expletive: ‘Fuck You!/Me!/Him!/It!’
D	Dest	Destinational usage: ‘Fuck off!’ ‘He fucked off’
E	EmphAdv	Emphatic adverb/adjective: ‘He fucking did it’ ‘in the fucking car’
F	Figurtv	Figurative extension of literal meaning: ‘to fuck about’
G	Gen	General expletive ‘(Oh) Fuck!’
I	Idiom	Idiomatic ‘set phrase’: ‘fuck all’ ‘give a fuck’
L	Literal	Literal usage denoting taboo referent: ‘We fucked’
M	Image	Imagery based on literal meaning: ‘kick shit out of’
N	PremNeg	Premodifying intensifying negative adjective: ‘the fucking idiot’
O	Pron	‘Pronominal’ form with undefined referent: ‘got shit to do’
P	Personal	Personal insult referring to defined entity: ‘You fuck!’/‘That fuck’
R	Reclaimed	‘Reclaimed’ usage—no negative intent, e.g. Niggers/Niggaz as used by African American rappers
T	Oath	Religious oath used for emphasis: ‘by God’
X	Unc	Unclassifiable due to insufficient context

Swearing rates in the full BNC datasets (Love, 2021)

Spoken BNC1994 (DS)

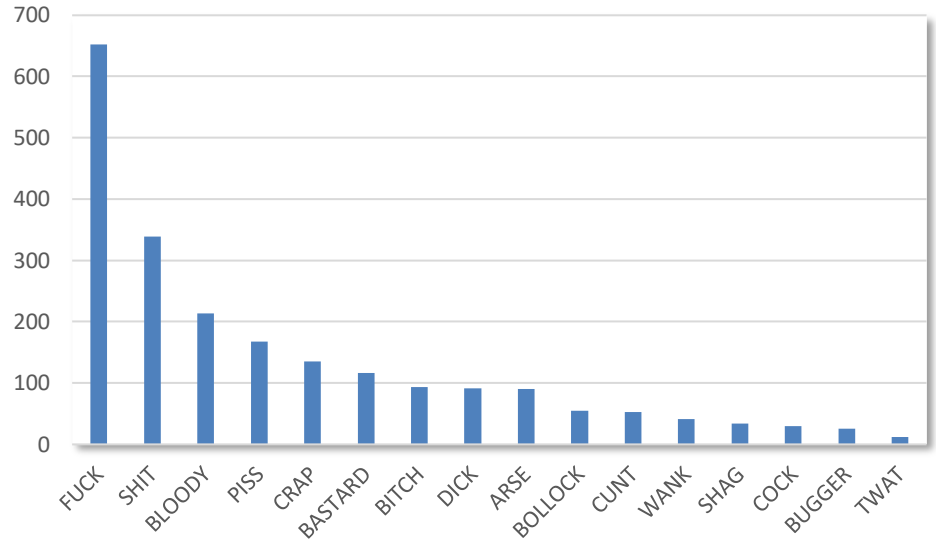


Spoken BNC2014

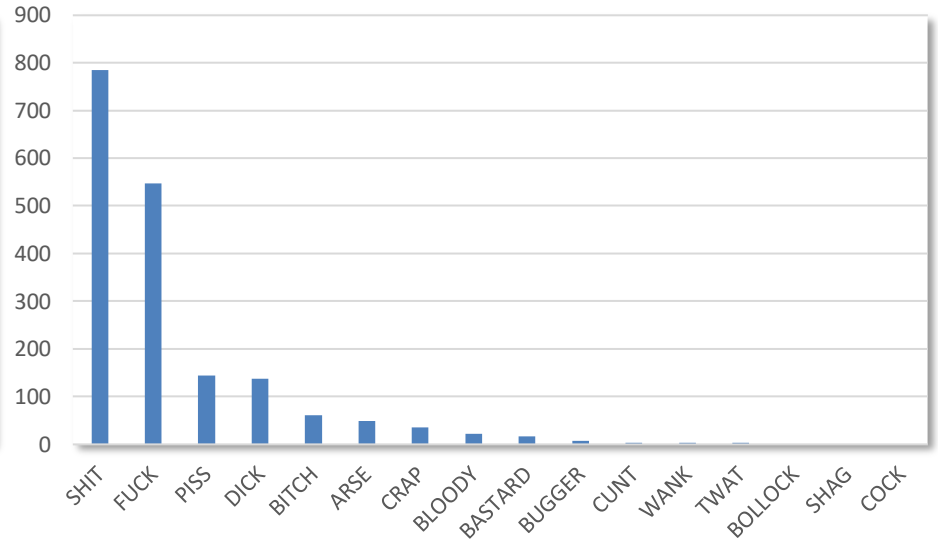


Swearing rates in COLT 1 and COLT 2

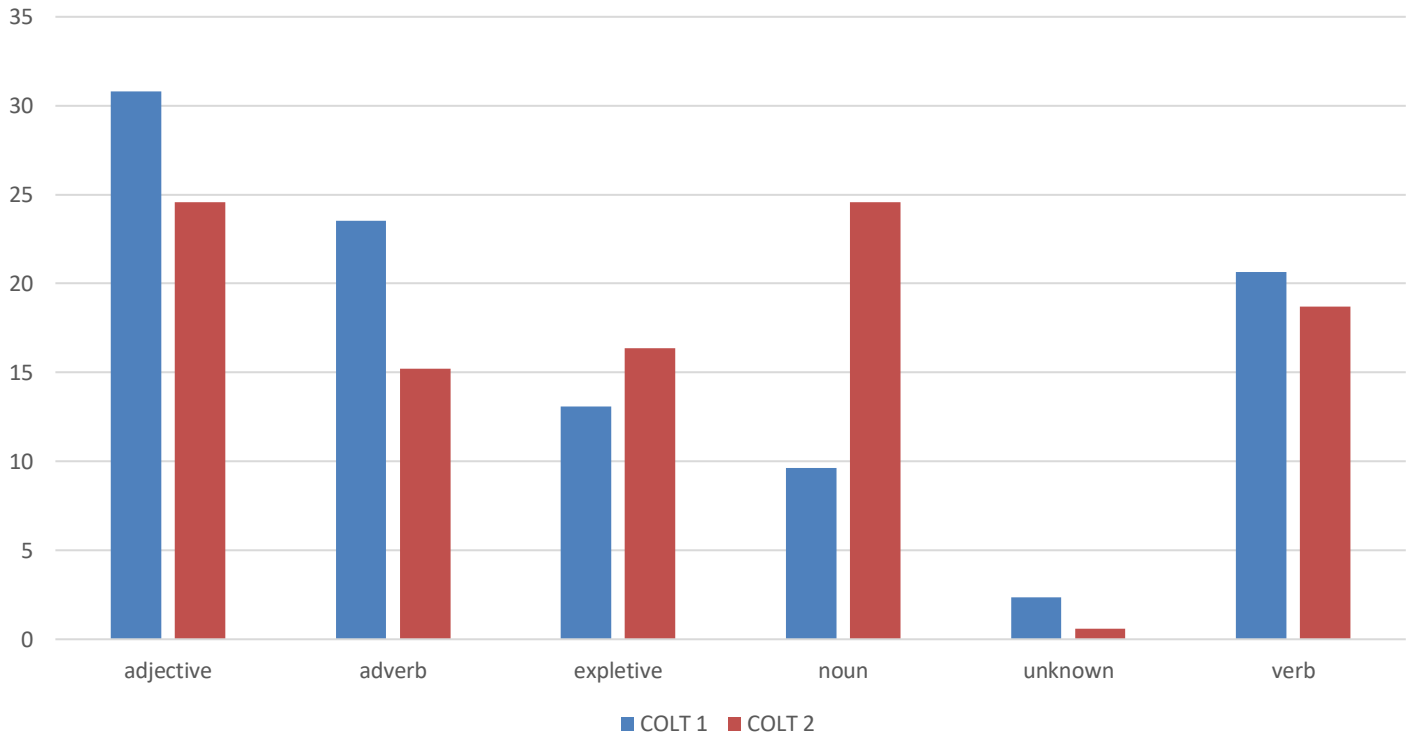
COLT 1



COLT 2



Word classes of FUCK



Functions of FUCK

COLT 1			COLT 2		
Code	**Description**	**%**	**Code**	**Description**	**%**
E	emphatic	39.5	E	emphatic	25.1
G	general expletive	12.4	I	idiomatic set phrase	22.2
B	adverbial booster	9.9	G	general expletive	16.4
D	destinational	8.1	F	figurative	12.3
I	idiomatic set phrase	6.1	A	predicative adjective	7.0

Functional comparison

Most common category in both is E: emphatic adverbs and adjectives

- *I can't be **fucking** bothered anymore* (COLT 1)
- *my teacher was a weird **fucking** psycho who fed us conspiracy theories* (COLT 2)

In COLT 2:

- more of category I: idiomatic set phrases (*fuck* as noun)
- e.g. *what/who/where the **fuck**, give a **fuck**, for **fuck's** sake, [ADJ] as **fuck**, shut the **fuck** up*
  - *we were just sitting here thinking okay **what the fuck** is going on?* (COLT 2)
  - *oh **for fuck's sake** when was that like four AM or something?* (COLT 2)

Functional comparison

In COLT 2:

- more of category G: general expletive
  - *I mean that's the reason you should want to go to uni oh yeah **fuck** yeah for the course if you if you end up it's not for the drinking* (COLT 2)

- more of category F: figurative extension of literal meaning (***fuck** up/over/about*)
  - *did I **fuck** something up?* (COLT 2)
  - *Jesus that's like two Jager Bombs each which yeah will **fuck** you over* (COLT 2)

Functional comparison

In COLT 2:

- less of category L: literal
  - *I know for definitely sure that Miss's **fucked** one of the upper sixth* (COLT 1)
  - *the beginning bit when she's **fucking** a man she sticks an axe through him* (COLT 1)

- Less of category D: destinational
  - *right then, I'll **fuck off** and see how you like it, eh?* (COLT 1)
  - ***fuck off** I'm not clearing the house* (COLT 1)

Reflections on swearing

The ranking of the top swear words in COLT 1 is reflected in the full Spoken BNC2014

- As shown in COLT 2, the rise in the popularity of SHIT seems to be driven by young speakers

Biggest development of FUCK is increased usage in idiomatic (fixed) expressions, e.g. *what the fuck*

- Also, there is a greater presence of figurative and general expletive usage…
  - …and general emphatic usage is (still) common
  - and there is less literal usage

- So, perhaps this points towards higher propensity for generalised usage in COLT 2

Reflections on ‘COLT 2’

- Retrofitting one corpus into another corpus design brings challenges
  - Unlike COLT 1, ‘COLT 2’ was not compiled as a separate entity and then added to the BNC
  - Instead, we are trying to extract an approximation of the COLT sampling frame from the BNC2014
  - This exercise necessarily involves compromise and the production of an imperfect match
- However...COLT 1 and COLT 2 do share three features:
  - Casual conversation
  - Among teenagers
  - From the south east of England
- We are just now exploring the utility of the comparison between COLT 1 and COLT 2
- Next steps: further data wrangling; functional analysis of swear word usage; evaluations of representativeness; sharing list of ‘COLT 2’ filenames

Aston University
BIRMINGHAM UK

Thank you

@lovermob
#CorpusCast

RiCL Research in Corpus Linguistics

About ▾ FirstView articles Current Archives Submissions Announcements

CFPs Special Issue of RiCL on "Innovations in the compilation and analysis of spoken corpora" edited by Robbie Love (Aston University)

References

BNC Consortium. (2007). *The British National Corpus, XML Edition*. Oxford Text Archive, http://hdl.handle.net/20.500.12024/2554

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics, 22(3), Special Issue: 'Compiling and analysing the Spoken British National Corpus 2014'*, 319-344.

Love, R. (2020). *Overcoming challenges in corpus construction: The Spoken British National Corpus 2014*. Routledge.

Love, R. (2021). Swearing in informal spoken English: 1990s – 2010s. *Text and Talk, 41, Special Issue: 'Corpus Linguistics across the Generations: In Memory of Geoffrey Leech'*.

McEnery, T. (2005). Swearing in English: Bad Language, Purity and Power from 1586 to the Present. Routledge.

McEnery, A., Baker, P. and Hardie, A. (2000) 'Swearing and Abuse in Modern British English', in B. Lewandowska-Tomaszczyk and P. Melia (eds) PALC'99: Practical Applications in Language Corpora, pp. 37–48. Berlin: Peter Lang

Stenstrom, A-B., Andersen, G., & Hasund, I. K. (2002). *Trends in Teenage Talk: Corpus compilation, analysis and findings*. Studies in Corpus Linguistics 8, John Benjamins.

Stenström, A-B. (2006). Taboo words in teenage talk: London and Madrid girls' conversations compared. *Spanish in Context, 3*(1), 115-138

Stenström, A-B., & Love, R. (in prep). London teenagers' use of fuck – now and then. *Journal of Pragmatics*.

Torgersen, E., Gabrielatos, C., Hoffmann, S. & Fox, S. (2011). A corpus-based study of pragmatic markers in London English. , 7(1), 93-118. https://doi.org/10.1515/cllt.2011.005