Aston University
BIRMINGHAM UK

Evaluating collocation in spoken dialogic corpora

Robbie Love, Aston University
Isobelle Clarke, Lancaster University
Mark McGlashan, Birmingham City University

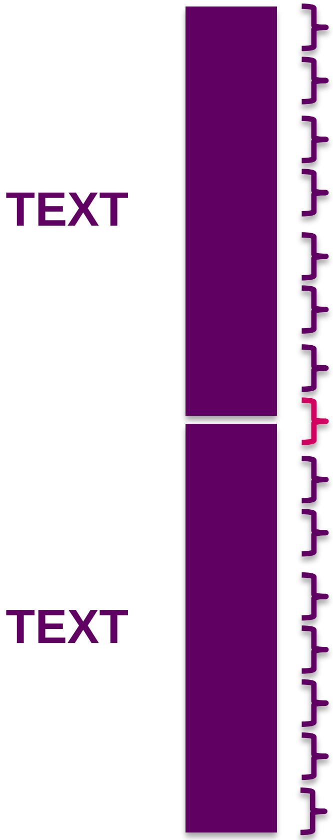
Context

- The study of collocation is a fundamental approach in the corpus linguistics toolkit
  - Firth (1957: 179) “You shall know a word by the company it keeps”

- Traditionally, three criteria for identifying collocations: **(i) distance**, (ii) frequency, and (iii) exclusivity (Brezina et al., 2015)
  - “distance specifies the span around a node word (the word we are interested in) where we look for collocates” (Brezina et al., 2015: 140)

- The ‘collocation window’ method measures collocation within a span, e.g. 5L & 5R (Gablasova et al., 2017: 158) – at 5L-5R, the tool searches for collocational patterns within strings of up to 11 tokens in length (five either side of the node, plus the node)
- This approach is facilitated by mainstream concordancers, allowing the user to define the collocational span according to their research interests

Context



TEXT

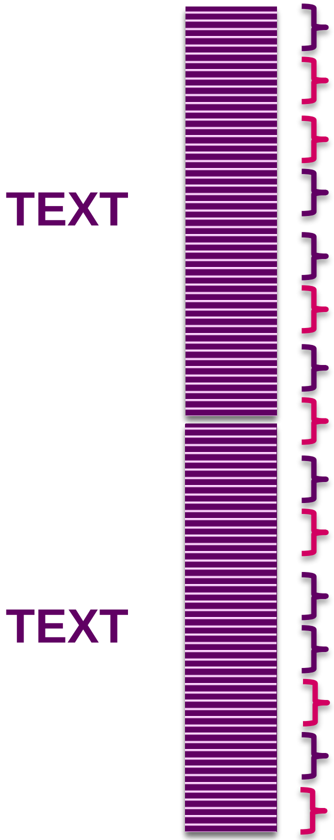
TEXT

- Many corpora have a **one-to-one correspondence between text and source** – all the material within an individual corpus text file comes from a single source
  - e.g. a corpus of news reporting, whereby each corpus text comprises a single news article

- With no restriction, as collocation of a given node is computed, most node-collocate pairs fall within a text

- Relatively few straddle across the boundary between the end of one text and the beginning of the next

Context



TEXT

TEXT

- Some corpora have a **one-to-many correspondence between text and source** – the material within an individual corpus text file comes from several sources
  - e.g. a spoken corpus containing dialogue between multiple speakers

- This means that, as well as text boundaries, there are internal boundaries between sources – in this case **utterance boundaries**

- With no restriction, as collocation of a given node is computed, there may be many node-collocate pairs that straddle utterance boundaries

The resulting collocation analysis may be based on a mixture of (a) collocate pairs produced by individual speakers and (b) collocate pairs ‘co-produced’ by two speakers

Collocation boundaries in the literature

“**[Collocates] may be in different sentences**, for example: *I wasn't altogether convinced by his argument. He had some strong points but they could all be met*. Clearly there are limits of relevance to be set to a collocational span of this but the question here is whether such limits can usefully be defined grammatically, and it is not easy to see how they can.”

(Halliday, 1966: 151-2)

“The notion of a purely linear collocational ‘span’, i.e. a stretch of a number of ‘orthographic words’ on either side, **disregarding sentence boundaries**, seemed to offer many theoretical as well as practical advantages; but the optimal solution consisted in ‘skipping’ certain ‘grammatical items’ which functioned merely as markers of a syntactic structure (rather than of a grammatical category).”

(Berry-Rogghe, 1970: 3)

On self-collocation: “**Very many […] are produced by the conversational situation itself**. There are some examples of question and answer: Is it *good*? / yes it’s *good* in a way”

(Jones & Sinclair, 1974: 46)

Collocation boundaries in the literature

“In other words, most of the lexical relations involving a word *w* can be retrieved by examining the neighborhood of *w*, wherever it occurs, within a span of five (-5 and +5 around *w*) words. In the work presented here, we use this simplification and consider that **two words co-occur if they are in a single sentence** and if there are fewer than five words between them.”

(Smadja, 1993: 151)

“Other decisions are whether to count only word tokens or all tokens (including punctuation and numbers), how to deal with multiword units (does out of count as a single token or as two tokens?), and **whether cooccurrences are allowed to cross sentence boundaries**.”

(Evert, 2008: 12)

“[L]aughter most often co-occurs with other features across the span of three turns and therefore **co-occurs across the boundaries of turns**, which means that it is used co-operatively or interactionally.”

(Schmidt, 2020: 216)

Collocation boundaries

- “Most studies on collocations do not take clause or sentence boundaries into consideration when specifying the collocation window” (Lehecka, 2015: 4)

- However, the facility to define collocation window boundaries when computing collocation is not something we had very often encountered or seen explicitly discussed in contemporary research

- This became evident when using the Spoken BNC2014 and noticing collocate pairs that straddled utterance boundaries, e.g.:

A: no I have I 've seen it

B: have we ?

Case study

Spoken BNC2014 (Love et al., 2017)

- c. 11 million words transcribed casual conversation
- 1,251 texts
- 672 speakers, L1 British English (2012-2016)

BNC
BRITISH
NATIONAL
CORPUS
2014

Access

- Pre-loaded as reference corpus in the following concordancers:
  - CQPweb (Hardie, 2012)
  - Sketch Engine (Kilgarriff et al., 2004)
  - #LancsBox (Brezina et al., 2021)

- Corpus file download: http://corpora.lancs.ac.uk/bnc2014/signup.php

Tools: facility to restrict by boundaries

Tools offering access to pre-loaded Spoken BNC2014

- CQPweb (Hardie, 2012) – no facility to restrict collocation window across boundaries
- Sketch Engine (Kilgarriff et al., 2004) – no facility to restrict collocation window across boundaries
- #LancsBox (Brezina et al., 2021) – facility to restrict collocation window across **sentence boundaries only**

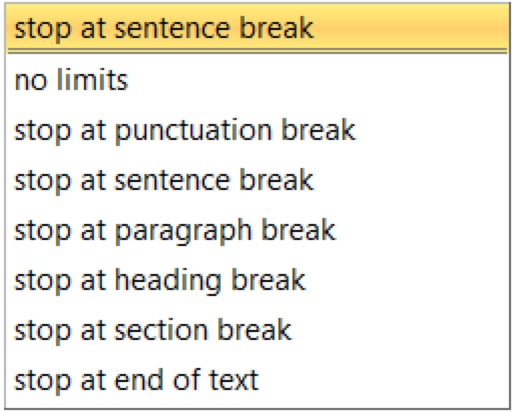
User upload of downloaded corpus files

- AntConc (Anthony, 2022) – no facility to restrict collocation window across boundaries
- WordSmith Tools (Scott, 2022) – facility to restrict collocation window across **various boundary types**

Procedure

- Corpus files stripped of XML markup
- ‘Heading’ inserted at start of each utterance (new line):

    <p>I 'm fed up of her getting up
    <p>mm she was n't too bad last night was she ?
    <p>no she slept



- Files uploaded to WordSmith Tools 8.0 (Scott, 2022)
- Collocation computed for selected node words, in the following conditions:
    - collocation boundary: *no limits* (= NO-BOUNDARY)
    - collocation boundary: *stop at heading break* i.e. utterance boundary (= U-BOUNDARY)
- Outputs compared between conditions – **what difference does it make to restrict by utterance?**

Procedure

NODE	RANK	FREQUENCY
I	#1	436,680
TO	#10	200,239
LIKE	#12	157,385
KNOW	#25	87,291
THINK	#42	54,465
CAN	#60	37,760
BEEN	#100	18,555
WEEK	#200	5,811
IDEA	#300	3,281
MAKING	#400	2,225

- Collocation computed for ten node words, range of wordlist frequencies (cf. Baker, 2016)
- Collocates below log-likelihood 15.13 ($p < 0.0001$) excluded
- Minimum collocate frequency: 5
- Collocates ranked by MI3 association measure
- Collocates compared for two conditions:
  - **NO-BOUNDARY**
  - **U-BOUNDARY**

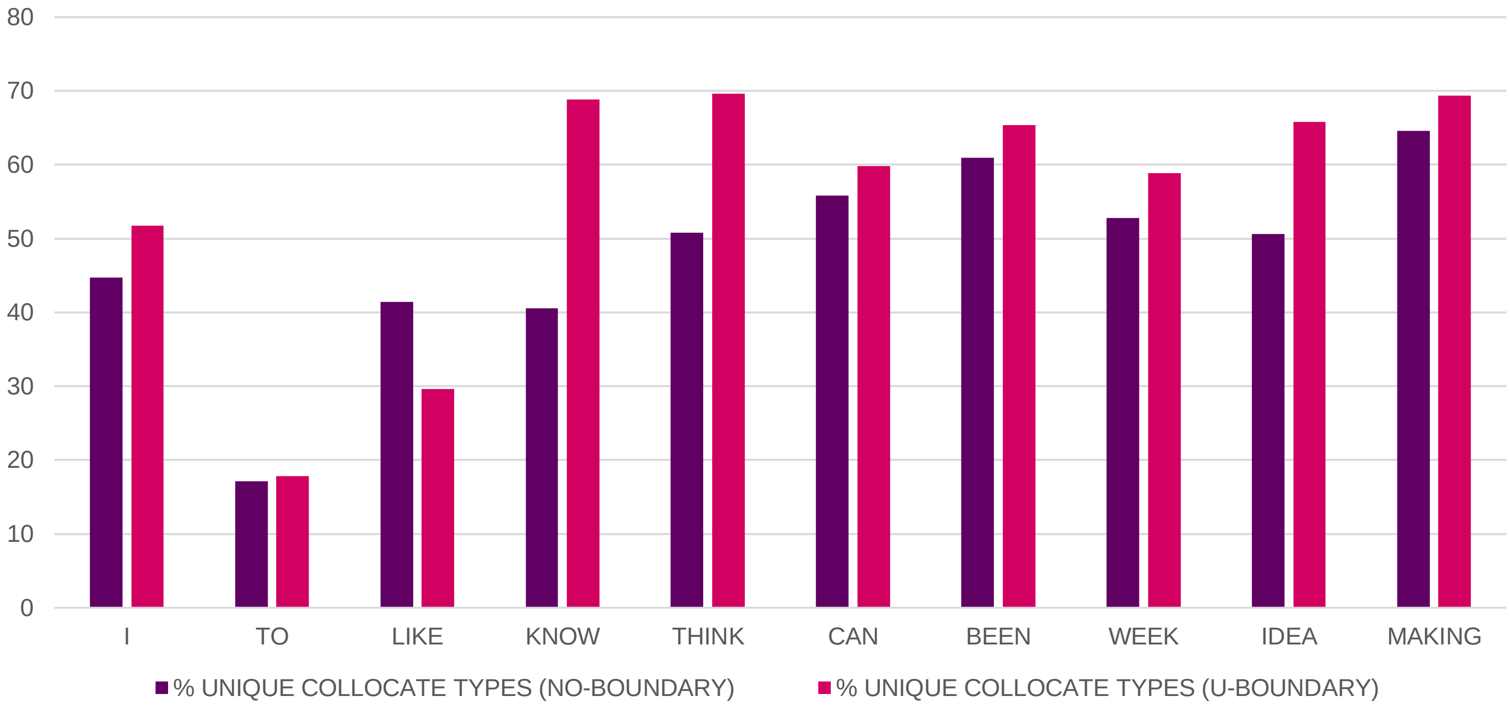
Findings: collocate token count

NODE	NO-BOUNDARY	U-BOUNDARY	% DIFF
I	3,439,310	2,651,486	-29.71
TO	1,273,004	1,500,989	15.19
LIKE	1,227,929	1,016,891	-20.75
KNOW	664,296	600,445	-10.63
THINK	342,881	295,010	-16.23
CAN	264,184	241,905	-9.21
BEEN	113,111	119,514	5.36
WEEK	32,392	31,360	-3.29
IDEA	17,787	17,575	-1.21
MAKING	7,751	8,879	12.70
		MEAN	**-5.78**

Findings: collocate type count

	NO-BOUNDARY		U-BOUNDARY	
NODE	TOTAL TYPES	UNIQUE TYPES	TOTAL TYPES	UNIQUE TYPES
I	1,891	846	2,166	1,121
TO	1,656	283	1,671	298
LIKE	1,041	431	867	257
KNOW	464	188	885	609
THINK	392	199	635	442
CAN	564	315	619	370
BEEN	351	214	396	259
WEEK	178	94	204	120
IDEA	81	41	117	77
MAKING	96	62	111	77

Findings: unique collocate types



Mean (NO-BOUNDARY) = **47.93%**

Mean (U-BOUNDARY) = **55.68%**

Findings: top collocates of *I*

	NO-BOUNDARY			U-BOUNDARY		
RANK	COLLOCATE	MI3	TOKENS	COLLOCATE	MI3	TOKENS
1	*N'T*	37.62	112,955	*N'T*	37.12	100,737
2	*DO*	37.12	92,394	*DO*	36.59	81,660
3	*IT*	36.82	118,199	*THINK*	36.5	61,417
4	*THINK*	36.68	65,362	*KNOW*	35.8	58,706
5	*YEAH*	36.39	96,540	*IT*	35.64	90,051
6	*KNOW*	36.28	65,640	*WAS*	35.22	58,563
7	*THAT*	35.99	84,638	*TO*	35.13	66,281
8	*AND*	35.98	89,634	*AND*	34.99	71,295
9	*TO*	35.83	77,903	*LIKE*	34.96	58,795
10	*WAS*	35.73	65,847	*THAT*	34.87	65,437
11	*YOU*	35.71	87,670	*VE*	34.62	38,539
12	*LIKE*	35.71	69,874	*THE*	34.5	65,358
13	*THE*	35.54	83,071	*YOU*	34.34	63,957
14	*VE*	34.95	41,621	*MEAN*	34.21	27,806
15	*BUT*	34.58	46,822	*JUST*	33.49	34,202
16	*NO*	34.57	45,213	*BUT*	33.44	36,044
17	*MEAN*	34.38	29,415	*YEAH*	33.37	48,218
18	*JUST*	34.16	39,913	*HAVE*	33.34	32,480
19	*SO*	34.06	41,623	*NO*	33.06	31,930
20	*HAVE*	33.98	37,596	*LL*	32.83	21,228

Findings: collocate types of *I*

- 846 (of 1,891) collocate types are **unique to NO-BOUNDARY condition**, i.e. they are **not** identified as collocates when restricting for utterance boundary, e.g. (top 10):

so, erm, er, at, ANONNAMEF, time, where, here, look, much

- 1,121 (of 2,116) collocate types are **unique to U-BOUNDARY condition**, i.e. they are **not** identified as collocates when no boundary restriction is in place, e.g. (top 10):

two, come, which, other, these, lot, into, way, stuff, little

  – e.g. *two*

NO-BOUNDARY	LL 5.77	MI3 27.44
U-BOUNDARY	LL 570.58	MI3 26.19

Findings: collocate frequency for *I*

- Ranking % difference in collocate frequency, i.e. the biggest reduction from NO-BOUNDARY → U-BOUNDARY

- These are words that are identified as collocates of *I* in both conditions, but have the biggest % difference in collocate frequency between the two conditions

- Several of these can be associated with turn-taking / interactional discourse marking – i.e. co-constructed collocation

RANK	COLLOCATE	NO-BOUNDARY	U-BOUNDARY	% DIFF
1	***HM***	574	150	-73.87
2	***MM***	27,692	8,285	-70.08
3	***YAY***	150	45	-70.00
4	***UHU***	836	255	-69.50
5	***DUH***	32	10	-68.75
6	***HMM***	503	172	-65.81
7	***YEP***	657	231	-64.84
8	***DEAR***	769	277	-63.98
9	***SEMI***	15	6	-60.00
10	***FALLS***	24	10	-58.33
11	***OPPOSED***	34	15	-55.88
12	***THANK***	1,415	636	-55.05
13	***FIELDS***	20	9	-55.00
14	***UNCLEARWORD***	30,117	13,785	-54.23
15	***ACADEMY***	13	6	-53.85
16	***OURSELVES***	41	19	-53.66
17	***COOL***	1,105	516	-53.30
18	***AH***	3,988	1,925	-51.73
19	***BRILLIANT***	513	255	-50.29
20	***YEAH***	96,540	48,218	-50.05

Findings: top collocates of *think*

	NO-BOUNDARY			U-BOUNDARY		
RANK	COLLOCATE	MI3	TOKENS	COLLOCATE	MI3	TOKENS
1	*THINK*	39.3	59,955	*THINK*	39.27	58,914
2	*IT*	32.16	20,202	*DO*	31.63	12,994
3	*DO*	32.06	14,375	*N'T*	30.78	11,667
4	*YOU*	31.59	16,959	*THAT*	29.74	10,019
5	*N'T*	31.26	13,028	*THE*	28.72	8,619
6	*THAT*	30.57	12,133	*TO*	28.07	6,488
7	*THE*	29.55	10,433	*WAS*	27.7	5,158
8	*THEY*	28.82	6,618	*AND*	27.55	6,405
9	*TO*	28.74	7,588	*OF*	27.52	4,982
10	*BUT*	28.72	6,051	*IS*	27.1	4,161
11	*AND*	28.72	8,392	*SO*	27.07	4,142
12	*SO*	28.2	5,377	*LIKE*	27	4,677
13	*OF*	28.18	5,797	*WE*	26.82	3,750
14	*HE*	28.12	5,170	*BE*	26.81	3,214
15	*IS*	28.02	5,147	*YEAH*	26.79	5,275
16	*LIKE*	27.91	5,774	*ABOUT*	26.47	2,461
17	*NO*	27.8	4,739	*WELL*	26.09	2,915
18	*JUST*	27.66	4,442	*WOULD*	25.92	2,157
19	*WELL*	27.48	4,017	*WHAT*	25.83	2,781
20	*BE*	27.46	3,735	*RE*	25.76	2,530

Discussion

- Restricting collocational measurement to speaker utterance does have effects on observations

- The extent of the effects is variable and requires further investigation
  - variability according to statistical measure, span size
  - effects on collocation networks (e.g. Brezina et al., 2015)
  - the role of visualisation

- While one effect is a net reduction in collocate frequency, another effect is variation in significant collocate types
  - **restricting to U-BOUNDARY does remove collocates, but it also introduces new ones**

Discussion

- Collocation across boundaries has been discussed in the literature. However, we argue that:
  - Relative to recent developments in the computation and visualisation of collocational relationships, collocational window boundaries appear to have been overlooked
  - Since collocational boundaries for utterances are not accounted for by popular concordancers, they are unlikely to be considered by many users

- When computing collocation, users working with spoken dialogic corpora should make (and report) an explicit decision on boundaries
  - While this is already possible (indirectly) in some concordancers, tool developers should consider introducing utterance boundary restriction as a feature
  - This will make the issue of collocational boundaries more visible and, therefore, something that more users are likely to take into account

Aston University
BIRMINGHAM UK

Thank you

r.love@aston.ac.uk	@lovermob
i.clarke@lancaster.ac.uk	@issy_clarke1
Mark.McGlashan@bcu.ac.uk	@Mark_McGlashan

References

Anthony, L. (2022). *AntConc (Version 4.1.3)* [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Baker, P. (2016). The shapes of collocation. *International Journal of Corpus Linguistics, 21*(2), 139-164. https://doi.org/10.1075/ijcl.21.2.01bak

Berry-Rogghe, G. (1970). *Collocations: Their computation and semantic significance*. PhD thesis, University of Manchester.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: a new perspective on collocation networks. *International Journal of Corpus Linguistics, 20*(2), 139-173. https://doi.org/10.1075/ijcl.20.2.01bre

Brezina, V., Weill-Tessier, P., & McEnery, A. (2021). *#LancsBox v. 6.0*. [software]. Available at: http://corpora.lancs.ac.uk/lancsbox

Evert, S. (2008). ‘Corpora and collocations’. In A. Lüdeling and M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*, article 58. Mouton de Gruyter, Berlin.

Firth, J. R. 1957. “A synopsis of linguistic theory 1930–1955”. In F. Palmer (Ed.), Selected Papers of J. R. Firth 1952–1959. London: Longman, 168–205.

Gablasova, D., Brezina, V., & McEnery, A. M. (2017). Collocations in corpus-based language learning research: identifying, comparing and interpreting the evidence. *Language Learning, 67*(Suppl. 1), 155-179. https://doi.org/10.1111/lang.12225

Halliday, M.A.K. (1966). ‘Lexis as a linguistic level’. In C. E. Bazell et al. (Eds), *In Memory of J. R. Firth*. Longman, 148–62.

Hardie, A. (2012). CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics, 17*(3), 380–409. https://doi.org/10.1075/ijcl.17.3.04har

Jones, S., & Sinclair, J. (1974). English lexical collocations: A study in computational linguistics. *Cahiers de lexicologie, 24*(1), 15-61. 10.15122/ISBN.978-2-8124-4277-3.P.0017

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., JMichelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography, 1*: 7-36. http://www.sketchengine.eu

Lehecka, T. (2015). ‘Collocation and colligation’. In J. Östman & J. Verschueren (Eds.), *Handbook of Pragmatics Online, 19*, John Benjamins (pp. 1-20).

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. DOI: 10.1075/ijcl.22.3.02lov

Scott, M., (2022). *WordSmith Tools version 8 (64 bit version)*. Stroud: Lexical Analysis Software.

Schmidt, S. (2020). *Rapport management in online spoken interaction: A cross-cultural linguistic analysis of communicative strategies*. PhD thesis, Birmingham City University.

Smadja, F. (1993). Retrieving Collocations from Text: Xtract. *Computational Linguistics, 19*(1), 143-177.