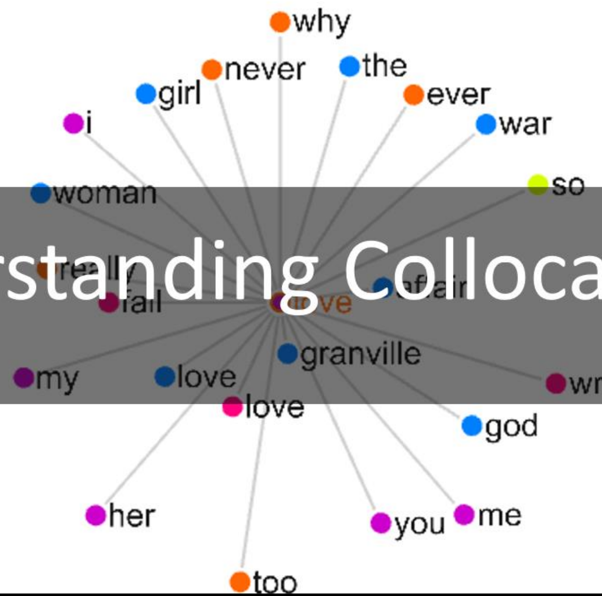
Corpus approaches to collocation: Challenges and opportunities

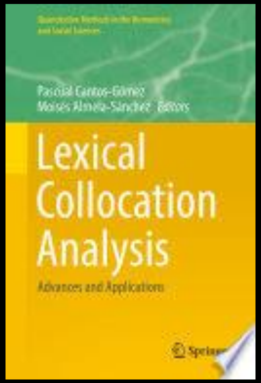
Dr. Vaclav Brezina & Dr. Dana Gablasova

Lancaster University



Understanding Collocations

> “
>
> In essence, collocation is a phenomenon concerned with repeated co-occurrence of words in texts. There is something profoundly simple, yet exceptionally insightful about the immediate space that words share with each other in texts (Brezina 2018:59).



Quantitative Methods in the Humanities and Social Sciences

Pascual Cantos-Gómez
Moisés Almela-Sánchez Editors

Lexical Collocation Analysis

Advances and Applications

Springer

Where to start?

What do we know about collocation?

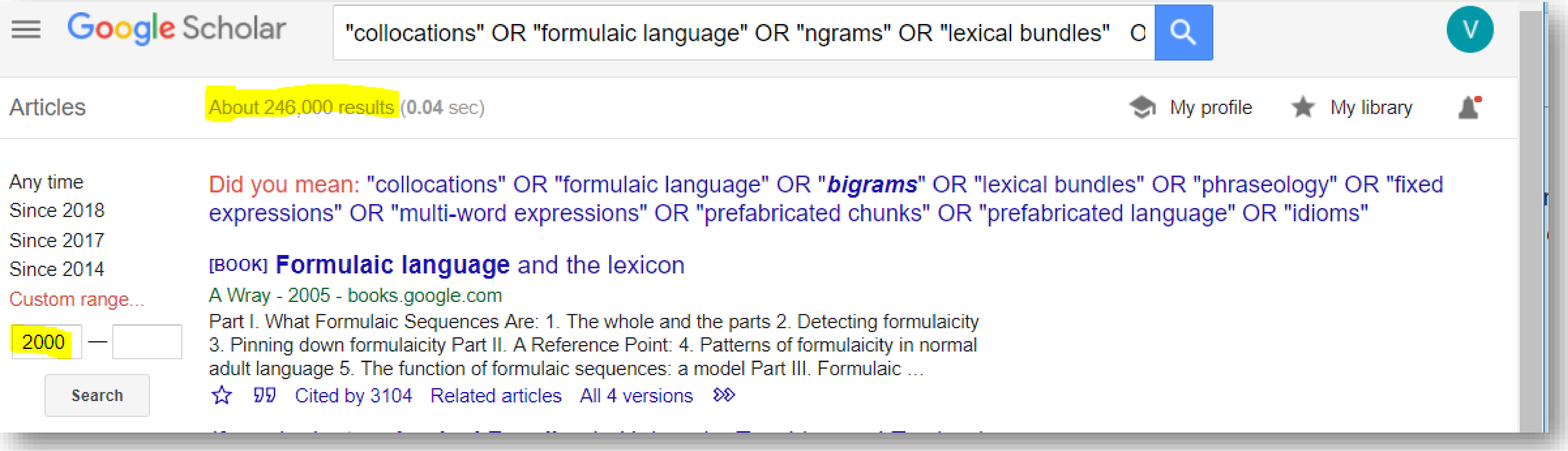
1. There is no consensus about the nature of the phenomenon.
2. There is no consensus about how to identify collocations.
3. There is no consensus about the terminology.

What do we know about ~~collocation?~~

1. There is no consensus about the nature of the phenomenon.
2. There is no consensus about how to identify collocations.
3. There is no consensus about the terminology.

What do we know about ~~collocation?~~

1. There is no consensus about the nature of the phenomenon.
2. There is no consensus about how to identify collocations.
3. There is no consensus about the terminology.
4. It is important.



"collocations" OR "formulaic language" OR "ngrams" OR "lexical bundles" O

Articles

About 246,000 results (0.04 sec)

Any time
Since 2018
Since 2017
Since 2014
Custom range...

2000 —

Search

Did you mean: "collocations" OR "formulaic language" OR "***bigrams***" OR "lexical bundles" OR "phraseology" OR "fixed expressions" OR "multi-word expressions" OR "prefabricated chunks" OR "prefabricated language" OR "idioms"

[BOOK] **Formulaic language** and the lexicon

A Wray - 2005 - books.google.com

Part I. What Formulaic Sequences Are: 1. The whole and the parts 2. Detecting formulaicity 3. Pinning down formulaicity Part II. A Reference Point: 4. Patterns of formulaicity in normal adult language 5. The function of formulaic sequences: a model Part III. Formulaic ...

Co-location as a starting point

amalgams – automatic – chunks – clichés – co-ordinate constructions – collocations – complex lexemes – composites – conventionalized forms – F[ixed] E[xpressions] including I[dioms] – fixed expressions – formulaic language – formulaic speech – formulas/formulae – fossilized forms – frozen metaphors – frozen phrases – gambits – gestalt – holistic – holophrases – idiomatic – idioms – irregular – lexical simplex – lexical(ized) phrases – lexicalized sentence stems – listemes – multiword items/units – multiword lexical phenomena – noncompositional – noncomputational – nonproductive – nonpropositional – petrifications – phrasemes – praxons – preassembled speech – precoded conventionalized routines – prefabricated routines and patterns – ready-made expressions – ready-made utterances – recurring utterances – rote – routine formulae – schemata – semipreconstructed phrases that constitute single choices – sentence builders – set phrases – stable and familiar expressions with specialized subsenses – stereotyped phrases – stereotypes – stock utterances – synthetic – unanalyzed chunks of speech – unanalyzed multiword chunks – units

Source: Wray (2002: 9)

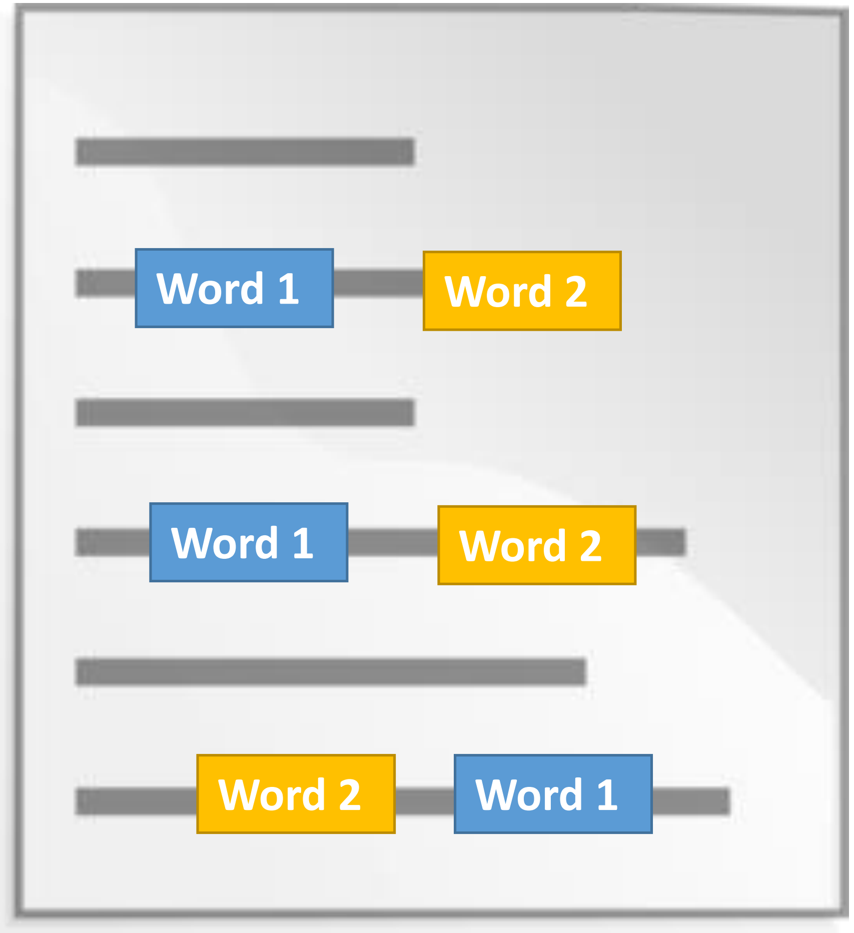
Co-location as a starting point

amalgams – automatic – chunks – clichés – co-ordinate constructions – **collocations** – complex lexemes – composites – conventionalized forms – F[ixed] E[xpressions] including I[dioms] – fixed expressions – formulaic language – formulaic speech – formulas/formulae – fossilized forms – frozen metaphors – frozen phrases – gambits – gestalt – holistic – holophrases – idiomatic – idioms – irregular – lexical simplex – lexical(ized) phrases – lexicalized sentence stems – listemes – multiword items/units – multiword lexical phenomena – noncompositional – noncomputational – nonproductive – nonpropositional – petrifications – phrasemes – praxons – preassembled speech – precoded conventionalized routines – prefabricated routines and patterns – ready-made expressions – ready-made utterances – recurring utterances – rote – routine formulae – schemata – semipreconstructed phrases that constitute single choices – sentence builders – set phrases – stable and familiar expressions with specialized subsenses – stereotyped phrases – stereotypes – stock utterances – synthetic – unanalyzed chunks of speech – unanalyzed multiword chunks – units

Source: Wray (2002: 9)

Co-location as a starting point (cont.)

- Low-inference category.
- Pre-theoretical.
- Data-driven.
- Close to textual reality.

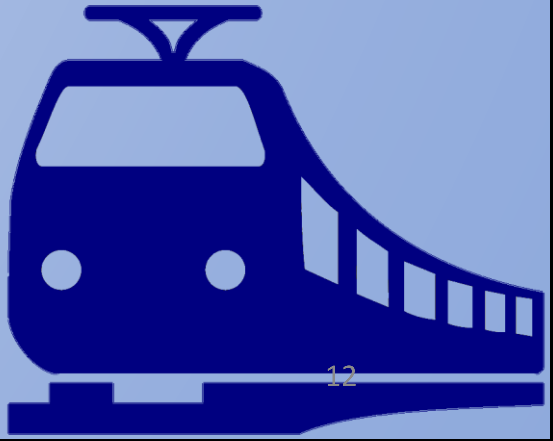


© David Parker

Source: Daily Mail

Reasons for co-location

1. Semantic unit (*carbon monoxide, global warming, okey dokey*).
2. Lexico-grammar (*of the, difference between)*.
3. Register preference (*large difference, administer a test, fucking stupid*).
4. Sociolinguistic choice (*sick movie, I believe*)
5. Discourse prosody (*illegal immigration, frail elderly*.)



Reasons for *studying* collocation

1. Language description (grammar, lexis, pragmatics etc.).
2. Discourse analysis (social, historical etc. meanings).
3. Language acquisition (L1 and L2).
4. Language pedagogy.
5. Language testing.

Corpus linguistics

Collocations

node

collocates

My **love** is like a red, red rose that’s newly sprung in June: My **love** is like the melody that’s sweetly played in tune. As fair art thou, my bonnie lass, so deep in **love** am I: And I will **love** thee still, my dear, till a’ the seas gang dry. Till a’ the seas gang dry, my dear, and the rocks melt wi’ the sun : And I will **love** thee still, my dear, while the sands o’ life shall run. And fare thee weel, my only **love**, and fare thee weel a while! And I will come again, my **love**, thou’ it were ten thousand mile.

(Robert Burns, “A Red, Red Rose”)

collocation window (span): 1L 1R

Random baseline model

My **love** is like a red, red rose that's newly sprung in June: My **love** is like the melody that's sweetly
played in tune. As fair a
will **love** thee still, my
dear, till a' the seas gar
ks melt wi' the sun : And I
will **love** thee still, my dear, while the sands o' life shall run. And fare thee weel, my only **love**, and
fare thee weel a while! And I will come again, my **love**, thou' it were ten thousand mile.

‘my love’ … 3

fare art And like red, sweetly in **love** **love**, And gang wi' played like dear, life shall rocks
sprung the Till deep my my And still, weel, again, ten the the while! is till And As I: a' only
come were sands sun:
ny bonnie My red is a
run. my **love** thee tho
dy thee a my am rose
love dear, that's **love** newly **love** fare **love**, will o' so dry. fair thee will that's in while June:
my seas tune. mile. thousand weel dear,

‘my love’ … 1

Association measures

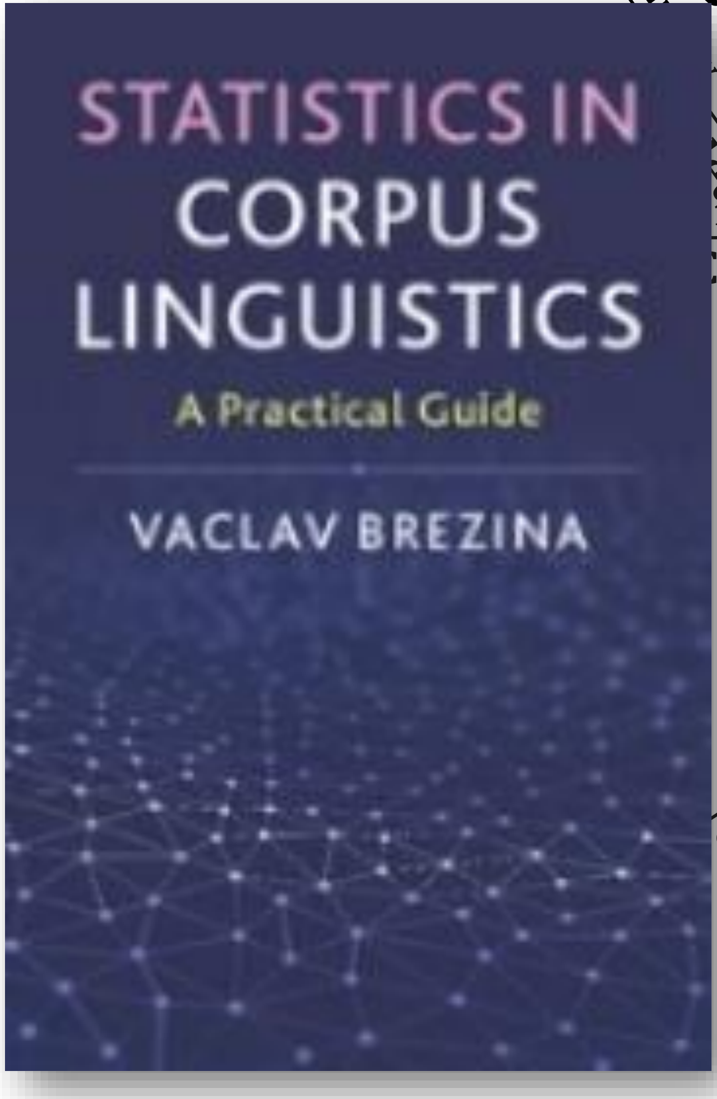
$$\log_2 \frac{O_{11}^{\;3}}{E_{11}}$$

$$O_{12} \times \log \frac{O_{12}}{E_{12}}$$

$$M_{in\ window} - M_{outside\ window}$$

pooled SD

$$C_1$$



STATISTICS IN CORPUS LINGUISTICS

A Practical Guide

VACLAV BREZINA

STATISTICS IN CORPUS LINGUISTICS

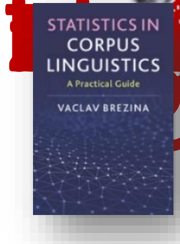
A Practical Guide

VACLAV BREZINA

$$\frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$$

$$\log_2 \frac{O_{11}^{\;2}}{E_{11}}$$

$$\log_2 \frac{O_{11} \times R_2}{O_{21} \times R_{1cor}}$$



STATISTICS IN CORPUS LINGUISTICS

A Practical Guide

VACLAV BREZINA

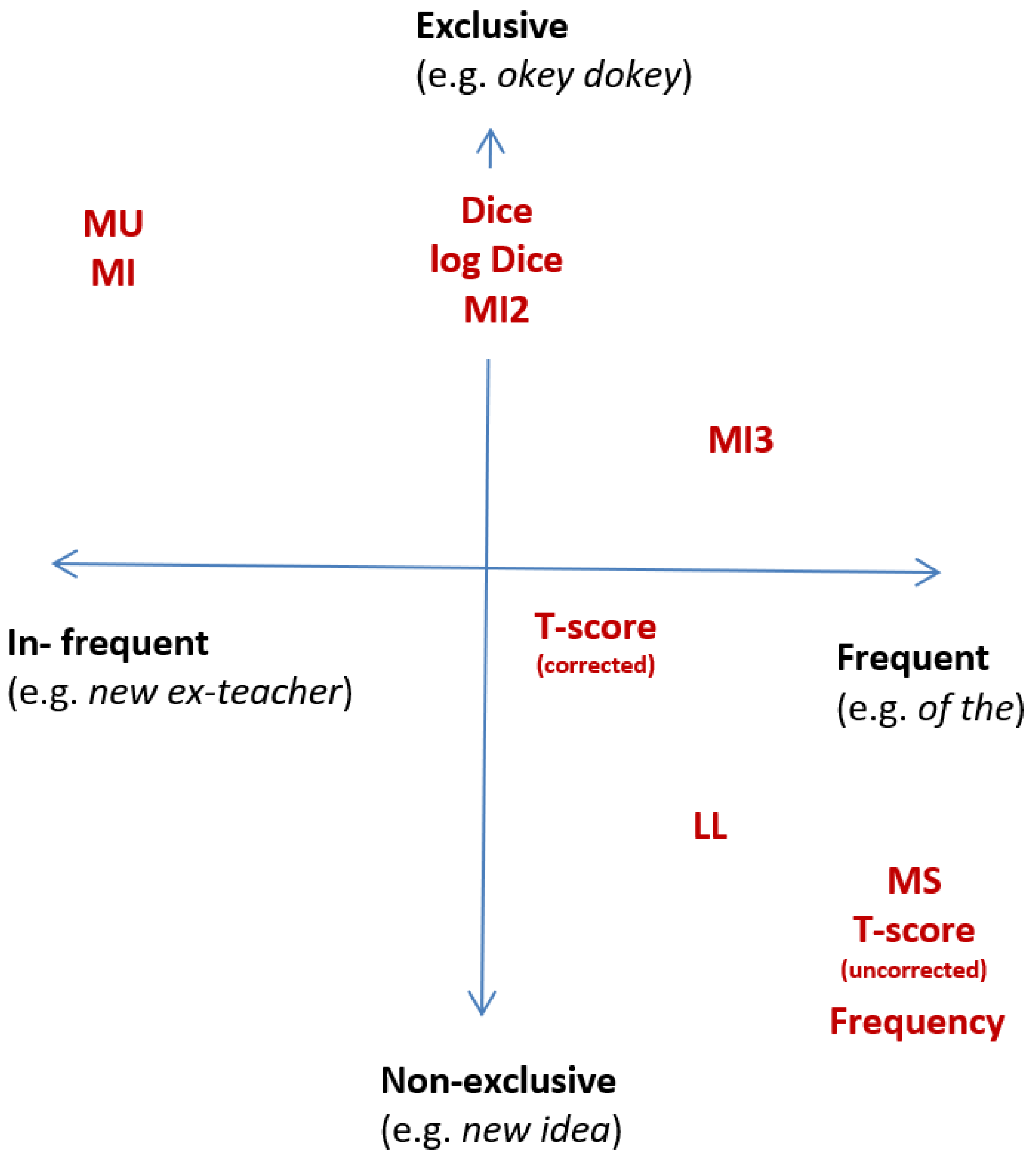
Log likelihood

Log Ratio

Cohen s d

$$\frac{O_{11}}{R_1} - \frac{O_{21}}{R_2}$$

Association measures (cont.)



Exclusive
(e.g. *okey dokey*)

MU
MI

Dice
log Dice
MI2

MI3

In- frequent
(e.g. *new ex-teacher*)

T-score
(corrected)

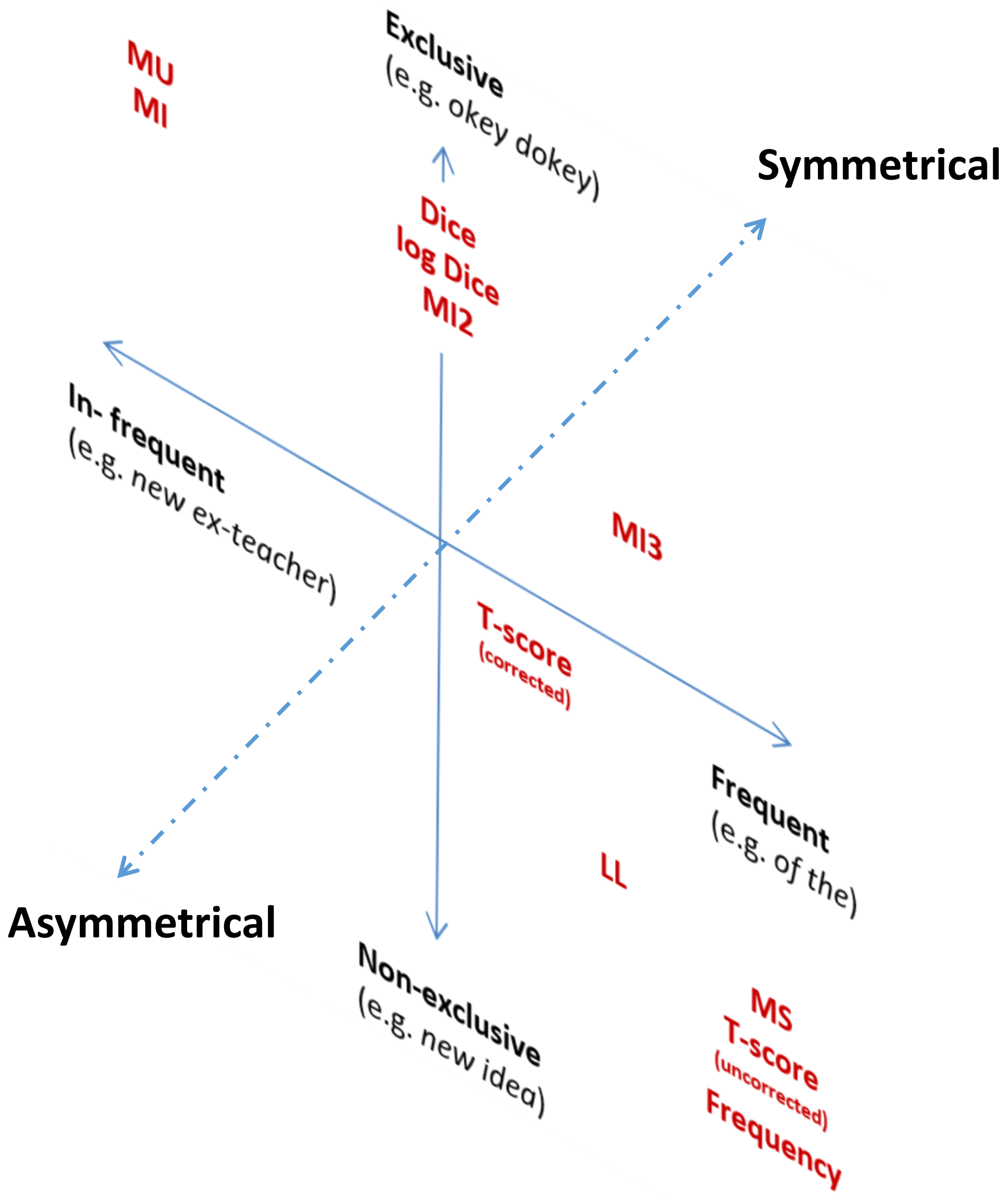
Frequent
(e.g. *of the*)

LL

MS
T-score
(uncorrected)
Frequency

Non-exclusive
(e.g. *new idea*)

Association measures (cont.)



Dimensions of collocation

1. Frequency of co-occurrence
   - Make a decision vs pay obeisance
2. Exclusivity
   - love affair ↔ love you
   - guinea pig, carbon monoxide
3. Directionality
   - extenuating → circumstances; circumstances → extenuating?
   - love → you; you → ?
4. Distance (span)
5. Connectivity (collocation networks)

In essence, collocation is a phenomenon concerned with repeated co-occurrence of words in texts. There is something profoundly simple, yet exceptionally insightful about the immediate space that words share with each other in texts .

BNC: 424 occurrences (3.78 per million words)

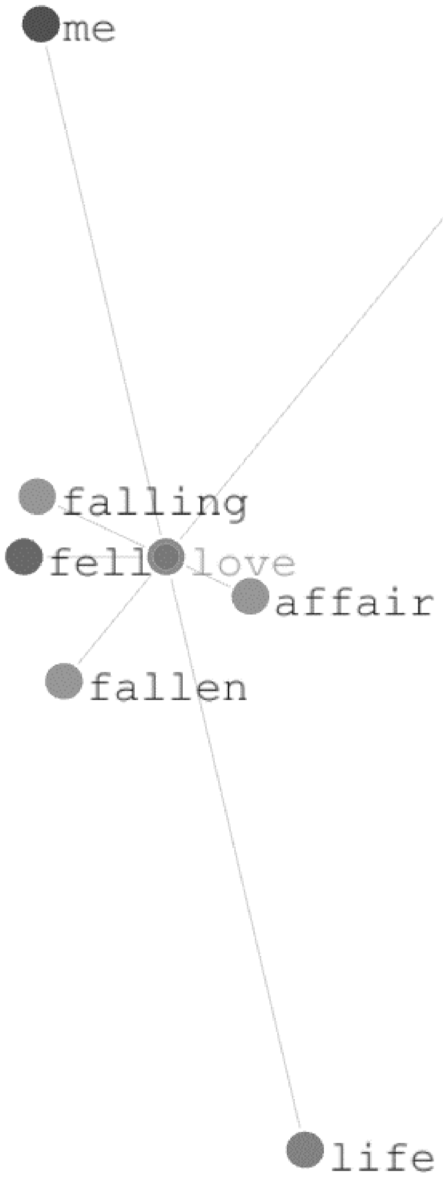
MU: 13.672	LL: 1,526.896 (p < 0.0001)	LOGDICE: 2.841
MI: 3.773	Z-score: 70.570	LOGRATIO: 4.137
MI2: 12.501	T-score: 19.085	MINIMUM SENSITIVITY: 0.000
MI3: 21.229	DICE: 0.000	DELTA P: [0.0002; 0.2191]

Visualizing collocations

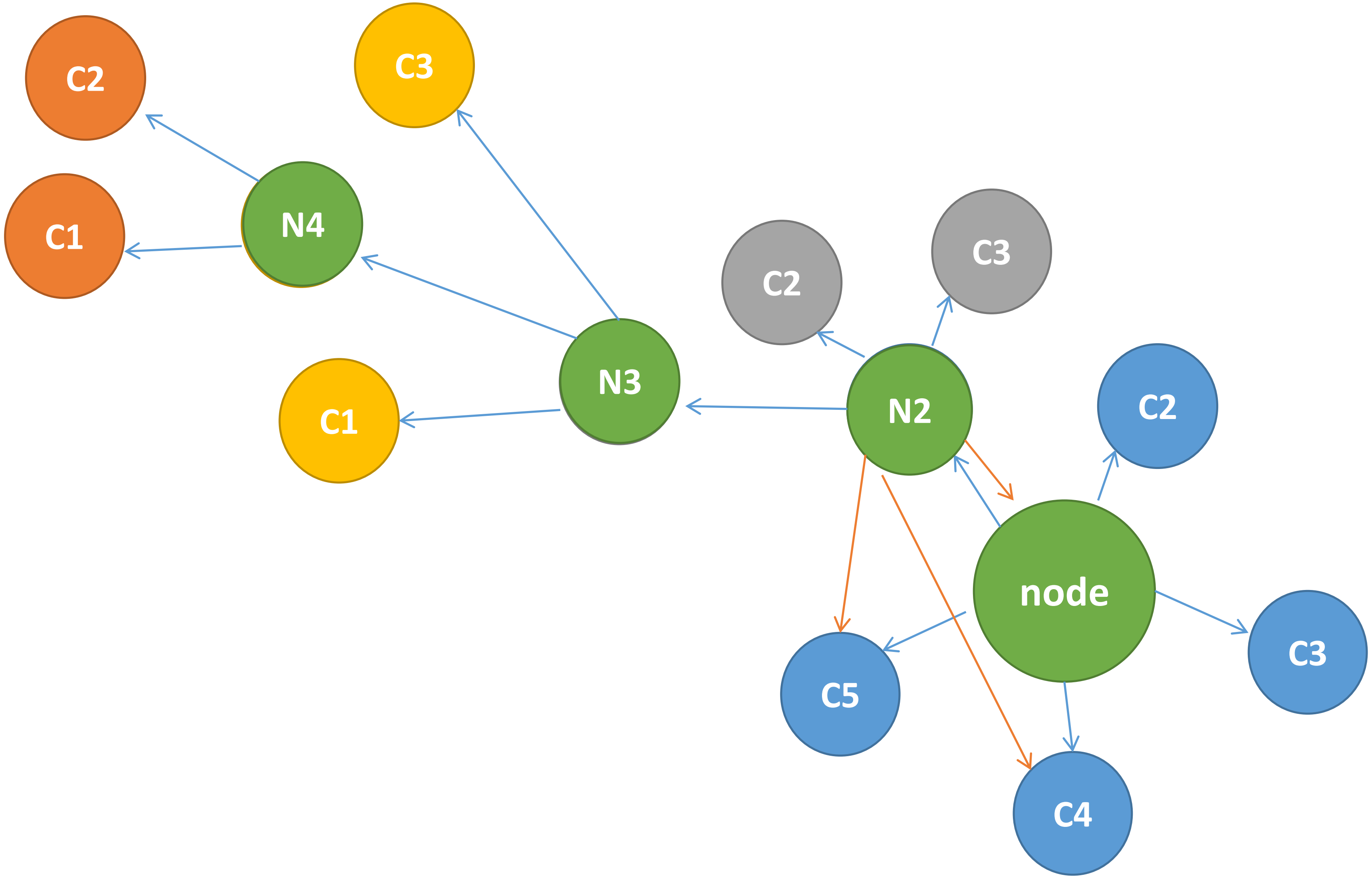
Traditional form of display

Collocate	MI-score	Freq (coll.)	Freq (corpus)
affair	8.86	5	37
fell	8.52	14	131
falling	8.52	5	47
fallen	8.37	5	52
me	5.57	23	1667
i'm	5.30	5	437
life	5.12	8	791

Collocation graph



Collocation networks

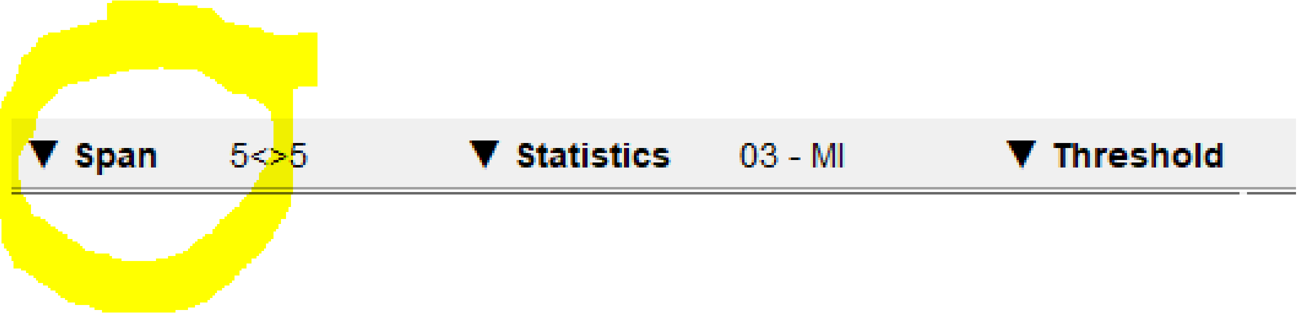


Everyday

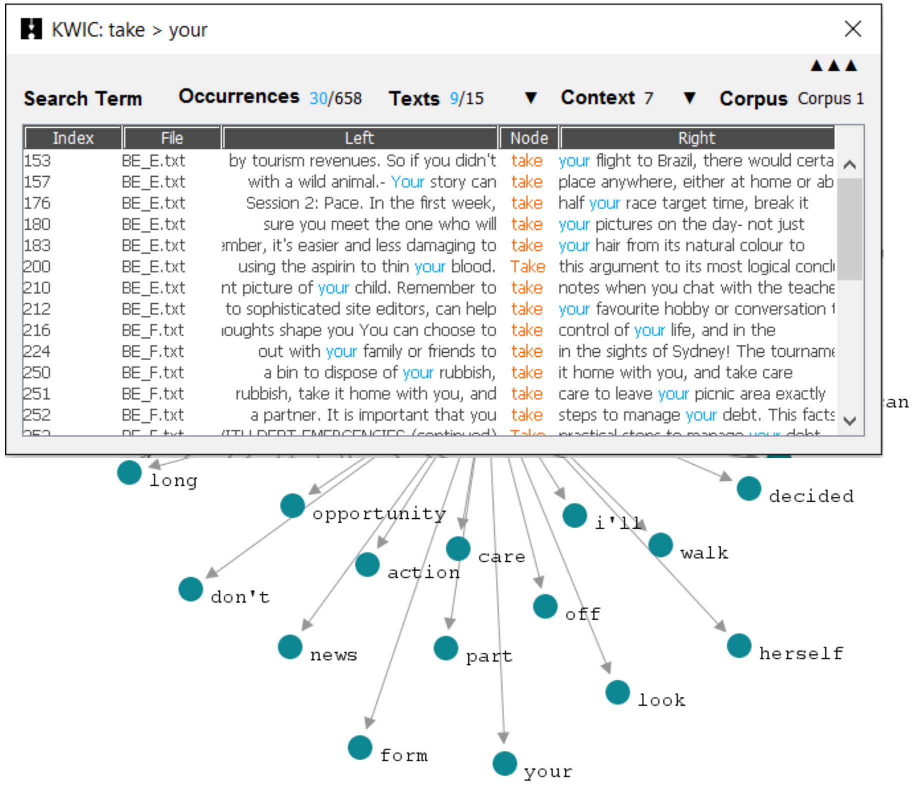
#LancsBox

Parameters

Span



▼ Span 5<>5 ▼ Statistics 03 - MI ▼ Threshold

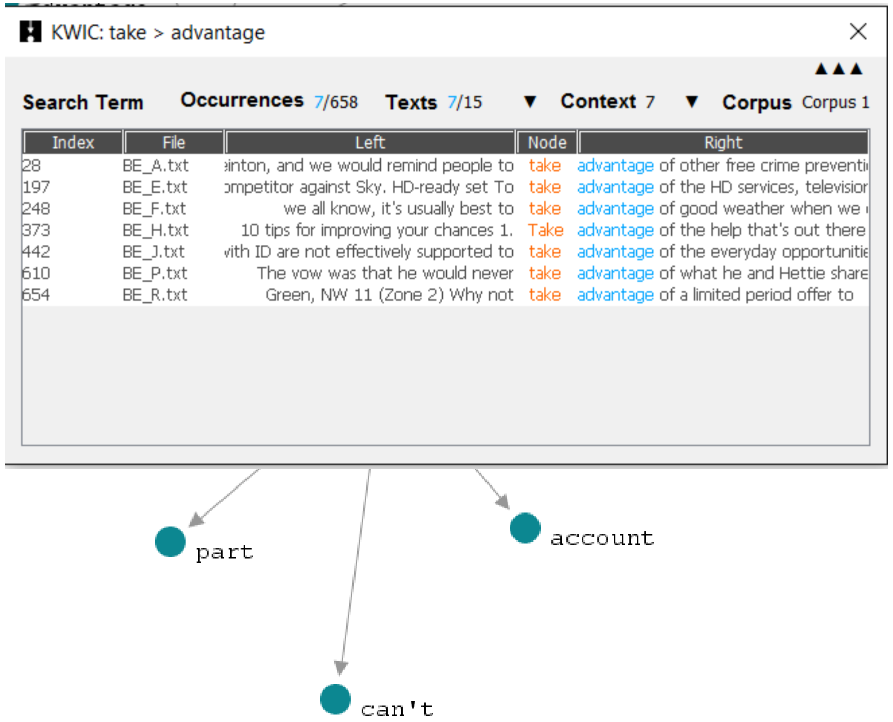


KWIC: take > your

Search Term Occurrences 30/658 Texts 9/15 ▼ Context 7 ▼ Corpus Corpus 1

Index	File	Left	Node	Right
153	BE_E.txt	by tourism revenues. So if you didn't	take	your flight to Brazil, there would certa
157	BE_E.txt	with a wild animal.- Your story can	take	place anywhere, either at home or ab
176	BE_E.txt	Session 2: Pace. In the first week,	take	half your race target time, break it
180	BE_E.txt	sure you meet the one who will	take	your pictures on the day- not just
183	BE_E.txt	ember, it's easier and less damaging to	take	your hair from its natural colour to
200	BE_E.txt	using the aspirin to thin your blood.	Take	this argument to its most logical concl
210	BE_E.txt	nt picture of your child. Remember to	take	notes when you chat with the teache
212	BE_E.txt	to sophisticated site editors, can help	take	your favourite hobby or conversation t
216	BE_F.txt	ioughts shape you You can choose to	take	control of your life, and in the
224	BE_F.txt	out with your family or friends to	take	in the sights of Sydney! The tourname
250	BE_F.txt	a bin to dispose of your rubbish,	take	it home with you, and take care
251	BE_F.txt	rubbish, take it home with you, and	take	care to leave your picnic area exactly
252	BE_F.txt	a partner. It is important that you	take	steps to manage your debt. This facts

5L, 5R: 36 collocates



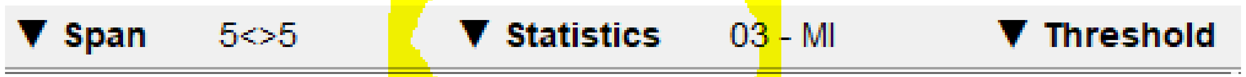
KWIC: take > advantage

Search Term Occurrences 7/658 Texts 7/15 ▼ Context 7 ▼ Corpus Corpus 1

Index	File	Left	Node	Right
28	BE_A.txt	einton, and we would remind people to	take	advantage of other free crime preventi
197	BE_E.txt	ompetitor against Sky. HD-ready set To	take	advantage of the HD services, televisior
248	BE_F.txt	we all know, it's usually best to	take	advantage of good weather when we
373	BE_H.txt	10 tips for improving your chances 1.	Take	advantage of the help that's out there
442	BE_J.txt	vith ID are not effectively supported to	take	advantage of the everyday opportunitie
610	BE_P.txt	The vow was that he would never	take	advantage of what he and Hettie share
654	BE_R.txt	Green, NW 11 (Zone 2) Why not	take	advantage of a limited period offer to

1L, 1R: 8 collocates

Statistic



▼ Span 5<>5 ▼ Statistics 03 - MI ▼ Threshold

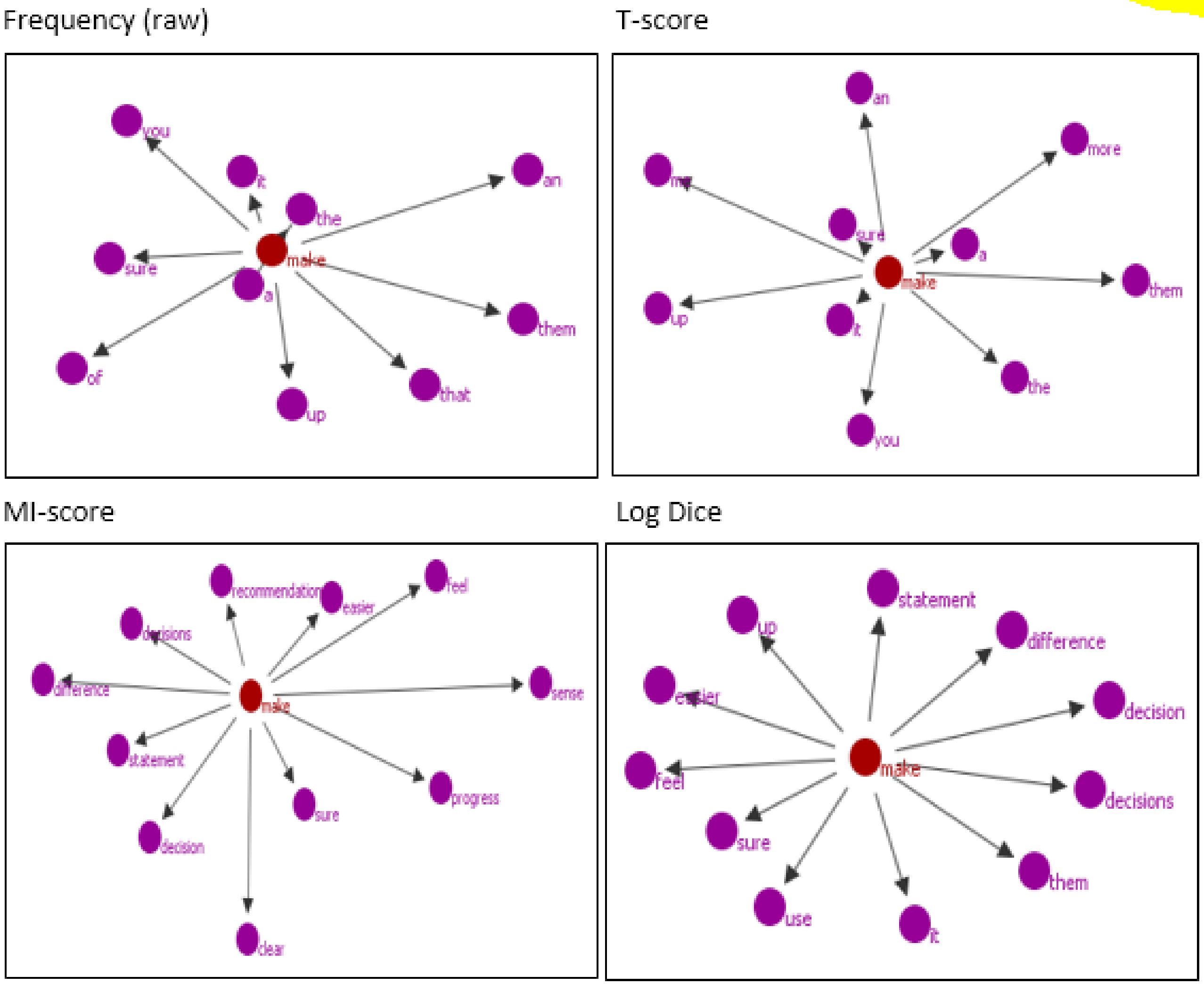
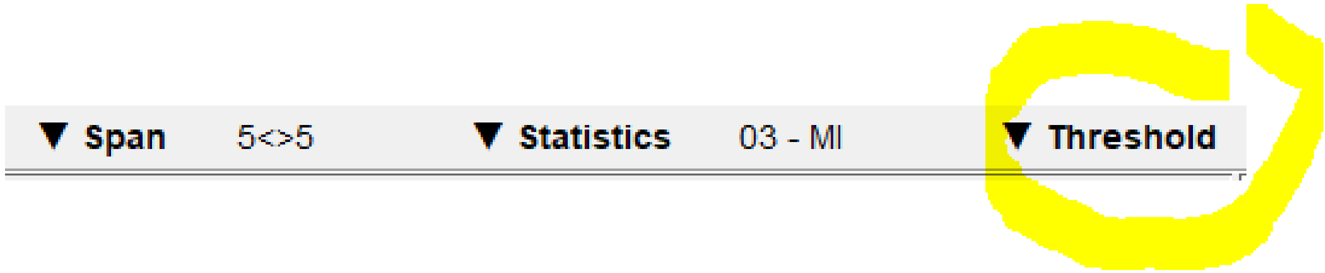


Figure 1 Top ten collocations of *make* for frequency and three AMs using L0, R2 windows in BE06 corpus

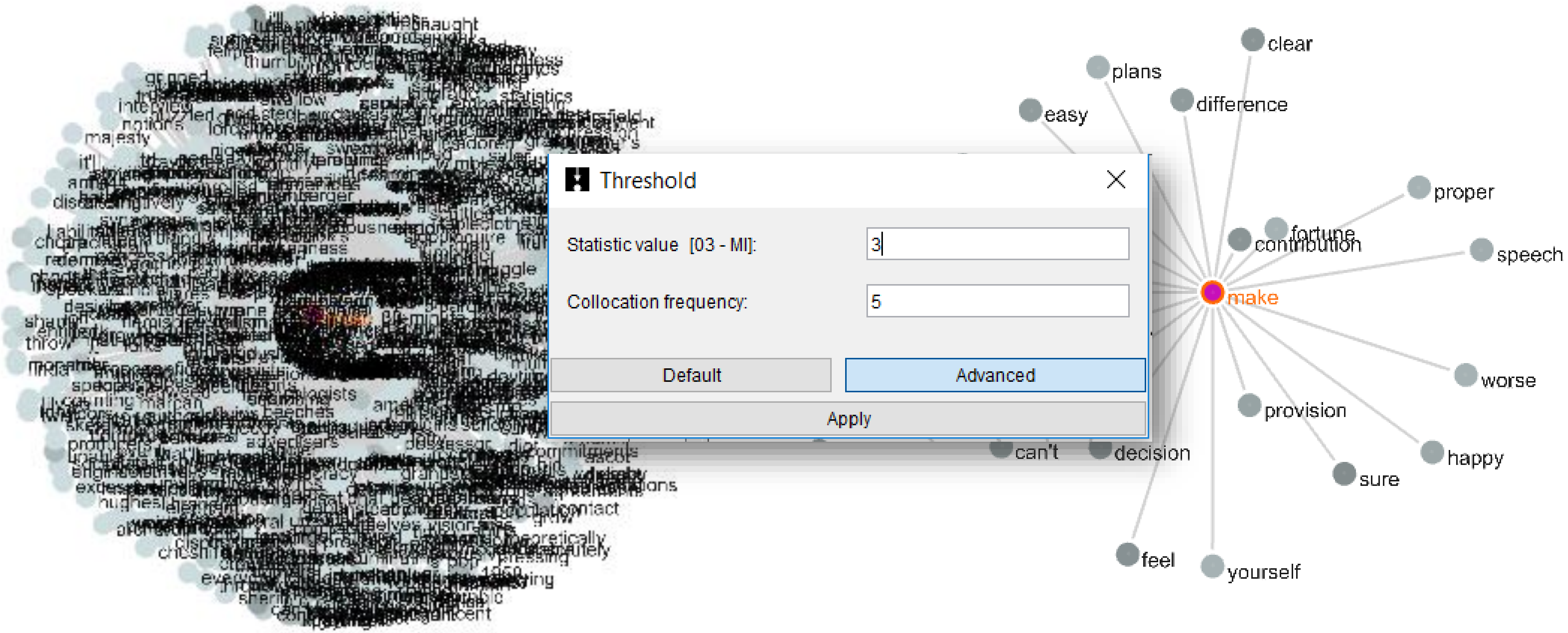
Threshold



▼ Span 5<>5

▼ Statistics 03 - MI

▼ Threshold



Threshold

Statistic value [03 - MI]: 3

Collocation frequency: 5

Low threshold

High threshold

CPN (Brezina et al. 2015)

Statistic ID	Statistic name	Statistic cut-off value	L and R span	Minimum collocate freq. (C)	Minimum collocation freq. (NC)	Filter
4b	MI2	3	L5-R5	5	1	function words removed
4b-MI2(3), L5-R5, C5-NC1; function words removed						

Example

Looking into the future

#LancsBox

#LancsBox and VR

#LancsBox: Experience corpora

#LancsBox is a new tool for corpus analysis and visualization developed at Lancaster University.

#LancsBox supports VR.

- Understanding the fabric of language.
- Experiencing language through corpora.
- Pedagogical applications.

Collocations in LLR

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, *67 (S1),* 155–179.

The state of play in language learning research

Use: Interest in frequency-based collocations (as part of formulaic language) **on the rise**; used to assess formulaic L2 production and compare it to L1 users

Method: Identifying collocations in the L2 production; deriving the AM values from a reference corpus (e.g. BNC); adding the values to the L2 production (e.g. Durrant & Schmitt, 2009) and compare to L1 use

Range of AMs: **limited**. Despite the existence of dozens of AMs - so far only a limited set used in LLR; **t-score and MI-score** dominant

Rationale for selection: AMs - **not fully understood** mathematical & linguistics procedures

> *"it is not clear which of these [MI-score and t-score] (or other) measures is the best to use in research, and to date, the selection of one or another seems to be somewhat arbitrary"* (González Fernández & Schmitt, 2015, p. 96)

The effect of register

Corpus	Size	Representativeness
British National Corpus (BNC)	98,560,118	Written and spoken (10M), diff. registers
BNC_Academic	15,778,043	Written, academic writing
BNC_News	9,412,245	Written, news
BNC_Fiction	16,143,913	Written, fiction
BNC – Context governed	6,196,134	Spoken, formal
BNC – Demographic	4,234,093	Spoken, informal

The effect of register (cont.)

make	BNC	Academic	News	Fiction	Formal speech	Informal speech
sure	6.8	7.09	7.26	5.78	6.9	6.64
decision	4.55	3.67	4.07	5.86	6.12	7.91
point	3.44	2.92	3.84	3.68	4.11	3.12

Replication?

Corpus	Size	Representativeness
BNC – Demographic (BNC_D)	4,234,093	Spoken, informal
BNC – 2014 Spoken (BNC_SP)	4,789,185	Spoken, informal
CANCODE (CANC)	5,076,313	Spoken, informal

Replication?

human	BNC_D	BNC_SP	CANC
beings	16.3	14.6	14.3
rights	12.2	11.6	9.4
nature	10.9	10.7	9.1

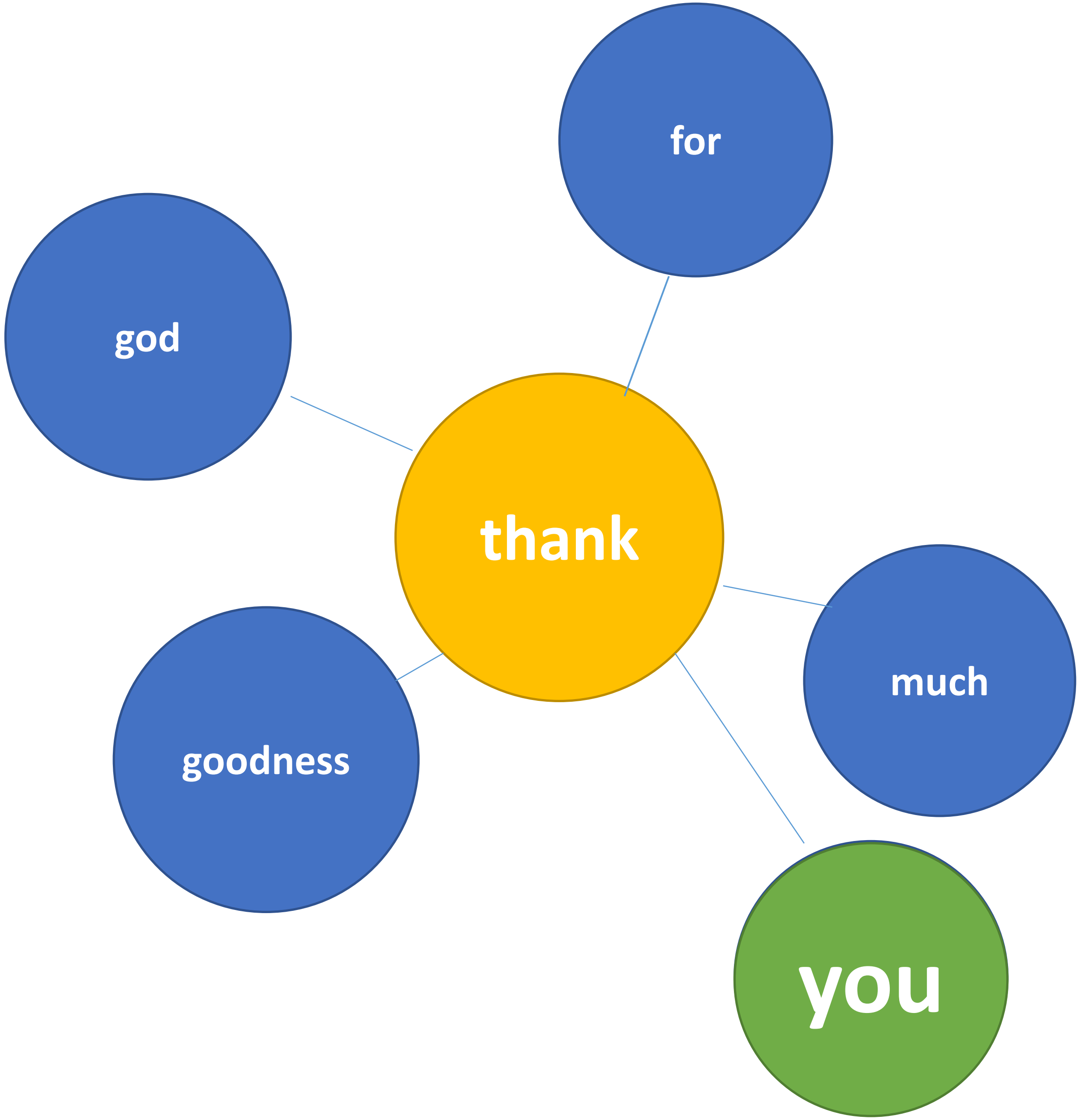
important	BNC_D	BNC_SP	CANC
vitally	14.36	13.62	11.28
terribly	8.39	-	7.28
very	6.22	5.33	6.03
really	2.79	3.86	3.54

To address the challenges in LLR

1. **Understand the AMs**: provide rationale for choice of measure, showing understanding of measure, why selected (beyond the fact that it was used by someone before)

2. **Consider a range of AMs and select an appropriate one** to reflect and capture the psycholinguistic concept that you hope to measure & suited to the specific RQ

3. **Consider the effect of genre and topic** (corpus representativeness) in interpretation of the L1 data



thank

for

god

much

goodness

you

References

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press.

Brezina, V. (2018). Collocation Graphs and Networks: Selected Applications. In *Lexical Collocation Analysis* (pp. 59-83). Springer.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, *67 (S1),* 155–179.

Brezina, V. (2016). Collocation Networks. In: Baker, P. & Egbert, J. (eds.) *Triangulating Methodological Approaches in Corpus Linguistic Research*. Routledge: London.

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.