

Lecture 3

A look at language collocations and
keywords

Overview

- The focus of this talk – the corpus linguistics perspective collocations and keywords
 - • Some examples
 - • Multi-methods

Frequency lists

- Simply a list of words and their frequencies in a corpus.

| | | |
|----|------|-------|
| 1 | the | 64420 |
| 2 | of | 31109 |
| 3 | and | 27002 |
| 4 | to | 26062 |
| 5 | a | 21972 |
| 6 | in | 18978 |
| 7 | is | 9961 |
| 8 | that | 9896 |
| 9 | was | 9368 |
| 10 | it | 9298 |

Collocation - Definition

collocations: the systematic co-occurrence of words in use
some examples: (you can probably think of your own):

- telephone - operator
- back – front (e.g. *back to front, front and back*)
- tell – story (e.g. *tell me a story*)

Collocates of diamond

multichip, solitaire, sapphire, jubilee, bingo, bracelet, graphite, brooch, cluster, necklace, waterfall, moth, ruby, presenter, ring, earrings, eternity, pearl, paste, engagement, shaped, tricks, rings, tiger, pin, anne, atoms, celebrated, dogs, rough, gold, dealer, shapes, mining, anniversary, victoria, neil, stone, queen, wedding

How large should the span be?

- Typically set at +/- 5 words. This seems to be the most useful span for collocates
- Similarly many people set a minimum threshold of frequency for words to count as collocates. I usually use a minimum frequency of 10
- Option to stop at sentence boundaries

Collocates of **company**

| | | |
|---|----------------|---|
| am a widow and he keeps the two of us | company | . Mrs Jackson, Grimston, Norfolk |
| to follow IBM slavishly) and also keeps the | company | on a smoother footing as it no |
| laugh, shows me affection constantly, keeps me | company | , gives me something to cuddle and |
| The bubbling cry of the curlew often keeps you | company | across the moorland to Laddow |
| the family, aren't you? Keeps me | company | .Say hello to Kevin.I wash it |
| under present arrangements;Either a | company | keeps going forward or it runs the |
| of Sir Arthur Sullivan, the D'Oyley Carte Opera | Company | keeps his music alive.Stars from |
| is covered in sanding marks, but in the | company | it helps us to keep the exception |
| .The haulier should make sure that his | company | secretary keeps a proper record of |
| borrowing.Credit Reference Agency: a | company | which keeps files on individuals' |
| translated: a word may be known by the | company | it keeps .One may look not only at |
| for Johnstone, the family name behind a | company | which keeps six members of the |
| been said that a person is judged by the | company | he keeps .It's also true that a |
| "Of him -- and the | company | he keeps .Go on." |

Rank by frequency

| No. | Word | Frequency | Frequency as collocate | In No. of texts |
|-----|-------------|-----------|------------------------|-----------------|
| 1 | <u>the</u> | 6,041,234 | <u>26151</u> | 2155 |
| 2 | , | 5,014,383 | <u>10307</u> | 1835 |
| 3 | <u>a</u> | 2,164,238 | <u>8314</u> | 1644 |
| 4 | <u>of</u> | 3,042,376 | <u>7039</u> | 1551 |
| 5 | . | 4,713,133 | <u>5969</u> | 1636 |
| 6 | <u>to</u> | 2,593,729 | <u>5553</u> | 1324 |
| 7 | <u>'s</u> | 783,990 | <u>4863</u> | 1184 |
| 8 | <u>and</u> | 2,616,708 | <u>4726</u> | 1433 |
| 9 | <u>in</u> | 1,937,819 | <u>4689</u> | 1438 |
| 10 | <u>that</u> | 1,118,985 | <u>2619</u> | 905 |

Mutual information

КОЛИЧЕСТВО ВЗАИМНОЙ информации

| No. | Word | Frequency | As collocate | In No. of texts | Mutual information value |
|-----|----------------------|-----------|--------------|-----------------|--------------------------|
| 1 | <u>ec-listed</u> | 6 | <u>5</u> | 1 | 8.6232085249112 |
| 2 | <u>anglo-persian</u> | 8 | <u>6</u> | 3 | 8.4712054314661 |
| 3 | <u>worshipful</u> | 49 | <u>34</u> | 24 | 8.3589959278801 |
| 4 | <u>anglo-iranian</u> | 13 | <u>9</u> | 4 | 8.3557282140462 |
| 5 | <u>clothworkers</u> | 17 | <u>11</u> | 6 | 8.2582117081319 |
| 6 | <u>'64</u> | 13 | <u>8</u> | 2 | 8.1858032126039 |
| 7 | <u>2/2</u> | 12 | <u>6</u> | 1 | 7.886242930745 |
| 8 | <u>beller</u> | 12 | <u>5</u> | 3 | 7.6232085249112 |
| 9 | <u>petrobras</u> | 17 | <u>7</u> | 7 | 7.6061350115522 |
| 10 | <u>d'oyly</u> | 35 | <u>13</u> | 9 | 7.4573996319411 |

Dice coefficient коэффициент Дайса

| No. | Word | Frequency | As collocate | In No. of texts | Dice coefficient |
|-----|---------------------------|-----------|----------------------|-----------------|------------------|
| 1 | insurance | 7,027 | 446 | 263 | 0.019 |
| 2 | parent | 3,728 | 311 | 151 | 0.0143 |
| 3 | shares | 8,386 | 340 | 128 | 0.0141 |
| 4 | record | 14,808 | 350 | 94 | 0.0128 |
| 5 | limited | 10,312 | 314 | 135 | 0.0125 |
| 6 | says | 39,194 | 482 | 223 | 0.0122 |
| 7 | 's | 783,990 | 4863 | 1184 | 0.0118 |
| 8 | holding | 7,963 | 261 | 121 | 0.0109 |
| 9 | has | 256,480 | 1608 | 625 | 0.0109 |
| 10 | private | 17,572 | 287 | 165 | 0.01 |

A look at language collocations and
keywords

Colligation

- A word collocates with a particular grammatical class.
- E.g 'he' colligates with verbs
- 'Mrs' colligates with proper nouns
- determiners colligate with nouns

Semantic preference

- Similar to Bill Louw's concept of semantic prosody.
- 'the relation, not between individual words, but between a lemma or word-form and a set of semantically related words' Stubbs (2001: 65)

Semantic preference – glass of

wine, sherry, champagne, beer, poured, water,
juice, brandy, milk, whisky, orange, lemonade,
rum, iced, sipped, gin, vodka, small, port,
cider, lager

Discourse prosody

- "Discourse prosodies express speaker attitude" (Stubbs 2001: 65)
- CAUSE (Stubbs 2001)
- General Corpus: CAUSE problem(s) 1806, damage 1519, death(s) 1109, disease 591, concern 598, cancer 572, pain 514, trouble 471
- Also: *cause* <accident, anger, chaos, crisis, doubt, hurt, suffering, upset>

Discourse prosody

- Environmental corpus: *cause <blindness, cancer, concern, damage, depletion, harm, loss, ozone, problem, radiation, warning>*

(Corpus) Keywords

A keyword list is calculated by comparing 2 frequency lists together – usually a much larger reference corpus against a smaller specialised corpus (but sometimes 2 equal sized corpora).

- Chi-square or log-likelihood test identify the words that are statistically much more frequent in one list when compared to the other.

<http://ucrel.lancs.ac.uk/llwizard.html>

Log-likelihood and effect size calculator

To use this wizard, type in frequencies for one word and the corpus sizes and press the calculate button.

| | Corpus 1 | Corpus 2 |
|-------------------|----------------------|----------------------|
| Frequency of word | <input type="text"/> | <input type="text"/> |
| Corpus size | <input type="text"/> | <input type="text"/> |

- Notes:
1. Please enter plain numbers without commas (or other non-numeric characters) as they will confuse the calculator!
 2. The LL wizard shows a plus or minus symbol before the log-likelihood value to indicate overuse or underuse respectively in corpus 1 relative to corpus 2.
 3. The log-likelihood value itself is always a positive number. However, my script compares relative frequencies between the two corpora in order to insert an indicator for '+' overuse and '-' underuse of corpus 1 relative to corpus 2.

How to calculate log likelihood

Log likelihood is calculated by constructing a contingency table as follows:

| | Corpus 1 | Corpus 2 | Total |
|--------------------------|----------|----------|---------|
| Frequency of word | a | b | a+b |
| Frequency of other words | c-a | d-b | c+d-a-b |
| Total | c | d | c+d |

Note that the value 'c' corresponds to the number of words in corpus one, and 'd' corresponds to the number of words in corpus two (N values). The values 'a' and 'b' are called the observed values (O), whereas we need to calculate the expected values (E) according to the following formula:

$$E_i = \frac{N_i \sum_j O_j}{n}$$

When is a word a keyword?

The analyst needs to apply cut-off points for statistical significance.

- Some analysts only look at the top 10 or 50 or 100 keywords instead.
- Additionally, sometimes a minimum frequency is applied (e.g. a word must occur 20 times before it's a keyword)
- Also, we may specify a keyword has to be reasonably well distributed (occurring in at least 20 texts)

Common types of keywords

- 1. Proper nouns (Clegg, Ghana etc)
- 2. Markers of style (often grammatical words like must, betwixt)
- 3. Spelling idiosyncrasies (color/colour)
- 4. “Aboutness” words (politics, recipe etc)

What's the point of it?

- Keywords identify salient words in a corpus, acting as signposts for a linguistic, cultural or discursive analysis. Explaining why they're there and what they do can lead to interesting and unexpected findings.
- Keywords can often not be predicted in advance as humans have cognitive biases when it comes to noticing frequencies.
- The statistical method is replicable and unbiased so it has a high reliability/validity from a scientific viewpoint.

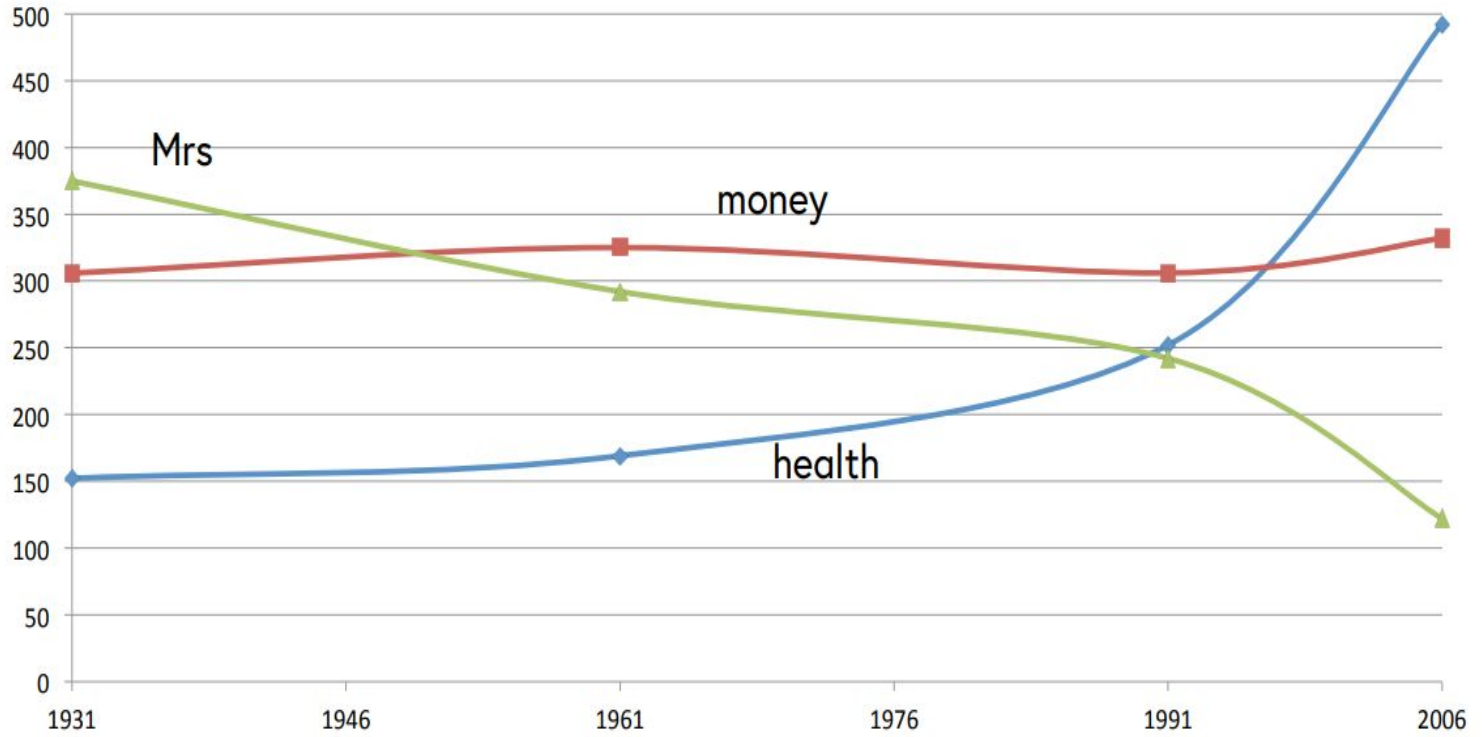
Example – Change over time (Baker 2011)

- How has British English changed over time? Which words have
 - a) steadily become less frequent
 - b) steadily become more frequent
 - c) stayed pretty much the same (lock words – locked in place)
- What can this tell us about cultural values?

Identifying key terms

- Four corpora used from 'The Brown Family' each a million words of written, published standard English (news, general prose, academic writing, fiction – 500 samples of 2000 words each), from 1931, 1961, 1991 and 2006.
- A technique called Coefficient of Variance used to identify words with highest and lowest changes in frequency over time.
- Only very frequent words (>1000 occurrences across all corpora considered)

Examples



Words that are declining the most

- Terms of address: *Mr, Mrs, sir* (a more informal society?)
- Stronger modal verbs: *shall, must, should* (a more democratic society?)
- Longer forms: *cannot, upon, half* (densification?)

Lock words

- Weaker modality: *can, could, would*
- Wh- words: *who, what, whether, where*
- Body parts: *hand, face, head, body, eyes*
- Other nouns: *life, world, government, money*

Words becoming more frequent

- Contracted forms: *it's, didn't, I'm, don't*
- Writing numbers as *34* rather than *thirty four*
- Social terms: *family, children, child, people, social, health, help*
- *Stable keywords vs. Emerging keywords*

...being a current smoker are elevated among children with intellectual disabilities (95 percent CE for
y . " People return to your homes ! Kill your children ! Kill your wives ! Then kill yourself ! " No
fective plans for safeguarding and promoting children 's welfare should be based on a wide-ranging ass
t the formal support for married women with children should be extended to single women and that un
nd dining table with chairs can not be taken . Children 's toys and other possessions can also not be tak
owing up I was probably one of thousands of children across the length and breadth of Wales whose da
them in a boarding school.⁹⁵ Many of these children were lucky enough to have relations who visited
: welfare of children and families Supporting children and families 1.1 All children deserve the opport
ts had a great influence on adults ' as well as children 's behaviour . We all went to confession , said o
f friends . In the same study , healthy-weight children were rated as clever , attractive , healthy , kind ,
opean Endeavour ; these were cards made by children visiting the National Maritime Museum Cornwa
torates published their report , Safeguarding Children , following joint inspections of children 's safeg
ed of offences listed under Schedule 1 of the Children and Young Persons Act 1933 would obviously c
ing . But Miss Connor stood up and told the children to leave their things exactly as they were on thei
es firmly on the road ahead and observed my children pinching each other in the back . If my wife was

Children

- In the top 10 most increasing words.
- In the 1931 corpus there are more references to "women and children" or "men, women and children".
- In the 2006 corpus there are references to moral panics around children: childhood obesity, paedophiles, poor literacy, poverty, knives, AIDS, adoption, protection from the media (e.g. advertising aimed at children), parental smoking and asthma – children are increasingly viewed as *victims and threats*.
- Does this fit the definition of a keyword as a socially prominent word with interlocking but sometimes contradictory meanings? Or is it too concrete a concept?

Multi Methods

- Corpora can answer some questions very well, others not at all.
- Corpora can integrate with other methods gainfully
- Corpora can help mesh quantitative and qualitative analyses
- Corpora are a tool – and like any tool they are good for some jobs and not others. They should also be part of a tool set.

Summing up

- Collocates and keywords are important techniques in corpus linguistics – you will come across the terms many times on this course
- They can tell us ‘about’ texts
- They can tell us about change over time
- They can help us decode argumentation strategies
- And more besides!

The

IMPOSSIBLE 2

QUIZ

GraphColl: Collocations in #LancsBox

- Collocation is systematic co-occurrence of words in text and discourse that we identify statistically

node

Love

From Wikipedia, the free encyclopedia

For other uses, see [Love \(disambiguation\)](#).

Love is a variety of different feelings, states, and attitudes that ranges from interpersonal affection ("I love my mother") to pleasure ("I loved that meal"). It can refer to an emotion of a strong attraction and personal attachment.^[1] It can also be a virtue representing human kindness, compassion, and affection—"the unselfish loyal and benevolent concern for the good of another".^[2] It may also describe compassionate and affectionate actions towards other humans, one's self or animals.^[3]

Ancient Greeks identified four forms of **love**: kinship or familiarity (in Greek, *storge*), friendship (*philia*), sexual and/or romantic desire (*eros*), and self-

collocation window (span): 2L 2R

variants of these states.^[7] This diversity of uses and meanings combined with the complexity of the feelings involved makes **love** unusually difficult to consistently define, compared to other emotional states.

Love in its various forms acts as a major facilitator of interpersonal relationships and, owing to its central psychological importance, is one of the most common themes in the creative arts.^[8]

Love may be understood as a function to keep human beings together against menaces and to facilitate the continuation of the species.^[9]

collocates

| | |
|----------------|----------------|
| from | agape |
| Wikipedia | modern |
| uses | romantic |
| see | non-Western |
| disambiguation | (2) traditions |
| love (2) | involved |
| is | makes |
| a | unusually |
| affection | difficult |
| I | in |
| my | its |
| mother | emotional |
| forms | states |
| of (2) | creative |
| kinship | arts |
| or (2) | may |
| divine | be |

BNC

Love: 22,265 hits in 1,983 different texts in 100M words

| Collocate | Frequency |
|-----------|-----------|
| of | 3,345 |
| i | 3,282 |
| and | 2,973 |
| in | 2,879 |
| to | 2,804 |
| the | 2,306 |
| you | 2,288 |
| with | 1,947 |
| for | 1,192 |
| a | 1,151 |

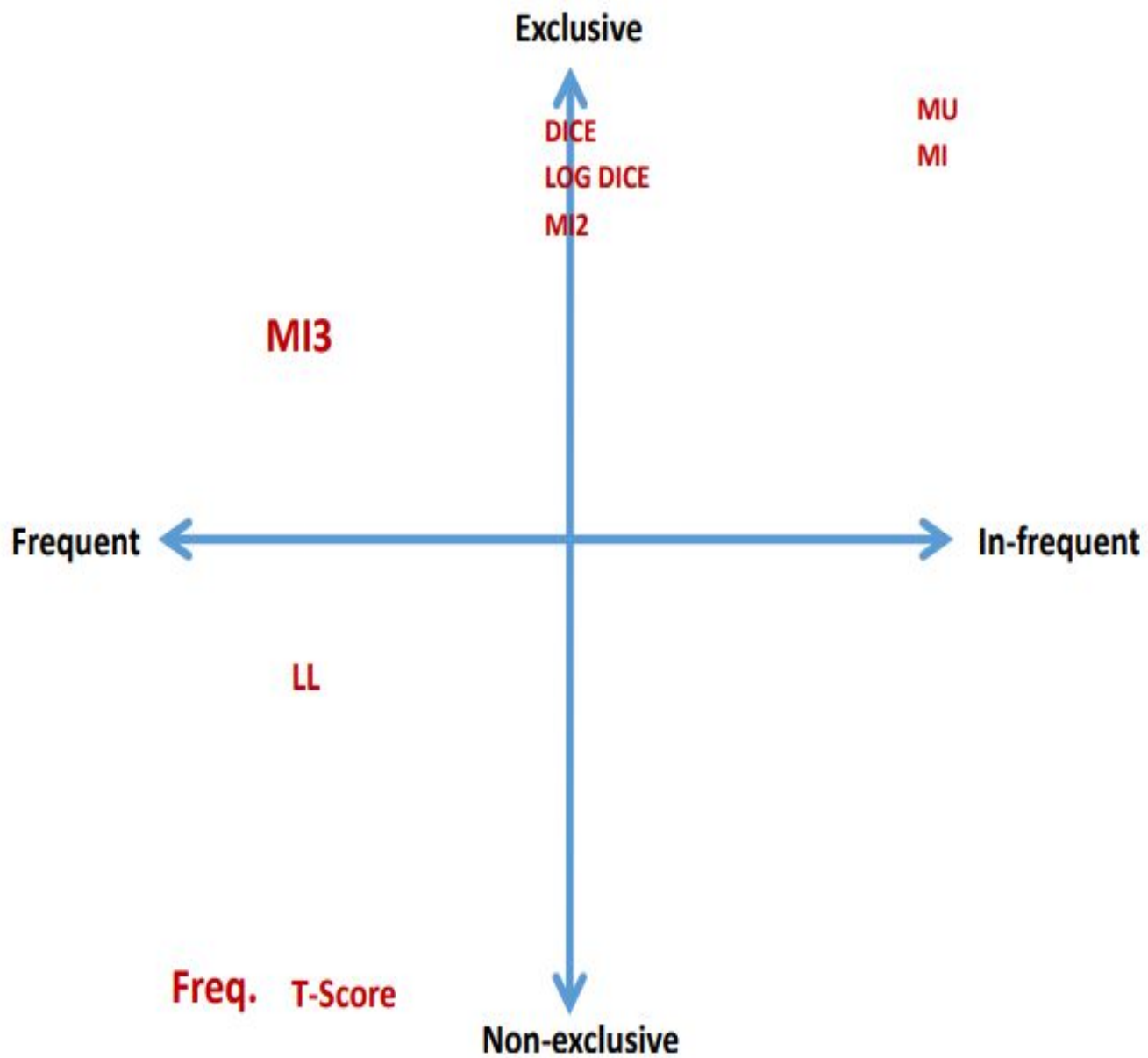
| Collocate | MI score |
|-----------|----------|
| affair | 6.3261 |
| fallen | 5.8542 |
| falling | 5.5687 |
| fell | 5.0409 |
| fall | 4.9721 |
| god | 4.4238 |
| making | 3.7614 |
| story | 3.5276 |
| 'd | 3.4879 |
| true | 3.3484 |

association
measure

Collocation

- Criteria:
- 1) distance -> span (e.g., L5, R5)
 - 2) frequency
 - 3) exclusivity
- } statistic, e.g. MI, MI3, LL, Log Dice...

| ID | Statistic | Equation | ID | Statistic | Equation |
|----|-------------------------|---|----|--------------------------|--|
| 1 | Freq. of co-occurrence | O_{11} | 8 | T-score | $\frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$ |
| 2 | MU | $\frac{O_{11}}{E_{11}}$ | 9 | DICE | $\frac{2 \times O_{11}}{R_1 + C_1}$ |
| 3 | MI (Mutual information) | $\log_2 \frac{O_{11}}{E_{11}}$ | 10 | LOG DICE | $14 + \log_2 \frac{2 \times O_{11}}{R_1 + C_1}$ |
| 4 | MI2 | $\log_2 \frac{O_{11}^2}{E_{11}}$ | 11 | LOG RATIO | $\log_2 \frac{O_{11} \times R_2}{O_{21} \times R_1}$ |
| 5 | MI3 | $\log_2 \frac{O_{11}^3}{E_{11}}$ | 12 | MS (Minimum sensitivity) | $\min\left(\frac{O_{11}}{C_1}, \frac{O_{11}}{R_1}\right)$ |
| 6 | LL (Log likelihood) | $2 \times \left(O_{11} \times \log \frac{O_{11}}{E_{11}} + O_{21} \times \log \frac{O_{21}}{E_{21}} + O_{12} \times \log \frac{O_{12}}{E_{12}} + O_{22} \times \log \frac{O_{22}}{E_{22}} \right)$ | 13 | DELTA P | $\frac{O_{11}}{R_1} - \frac{O_{21}}{R_2}; \frac{O_{11}}{C_1} - \frac{O_{12}}{C_2}$ |
| 7 | Z-score | $\frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$ | 14 | Cohen's d | $\frac{Mean_{in\ window} - Mean_{outside\ window}}{pooled\ SD}$ |



Collocation (cont.)

Criteria:

1) distance -> span (e.g., L5, R5)

2) frequency

3) exclusivity

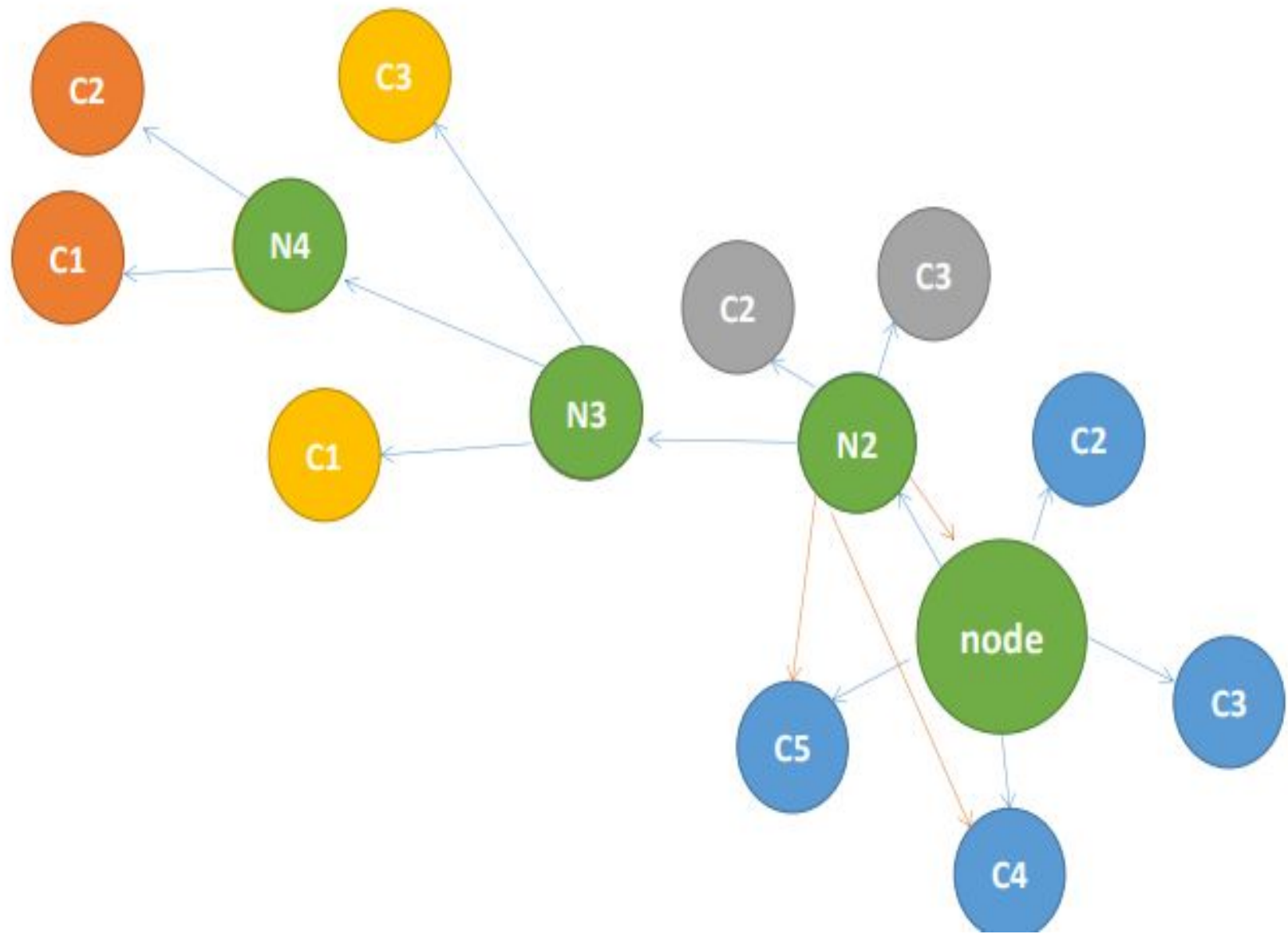
} statistic, e.g. MI, MI3, LL, Log Dice...

4) dispersion

5) directionality: Delta P

6) type-token distribution

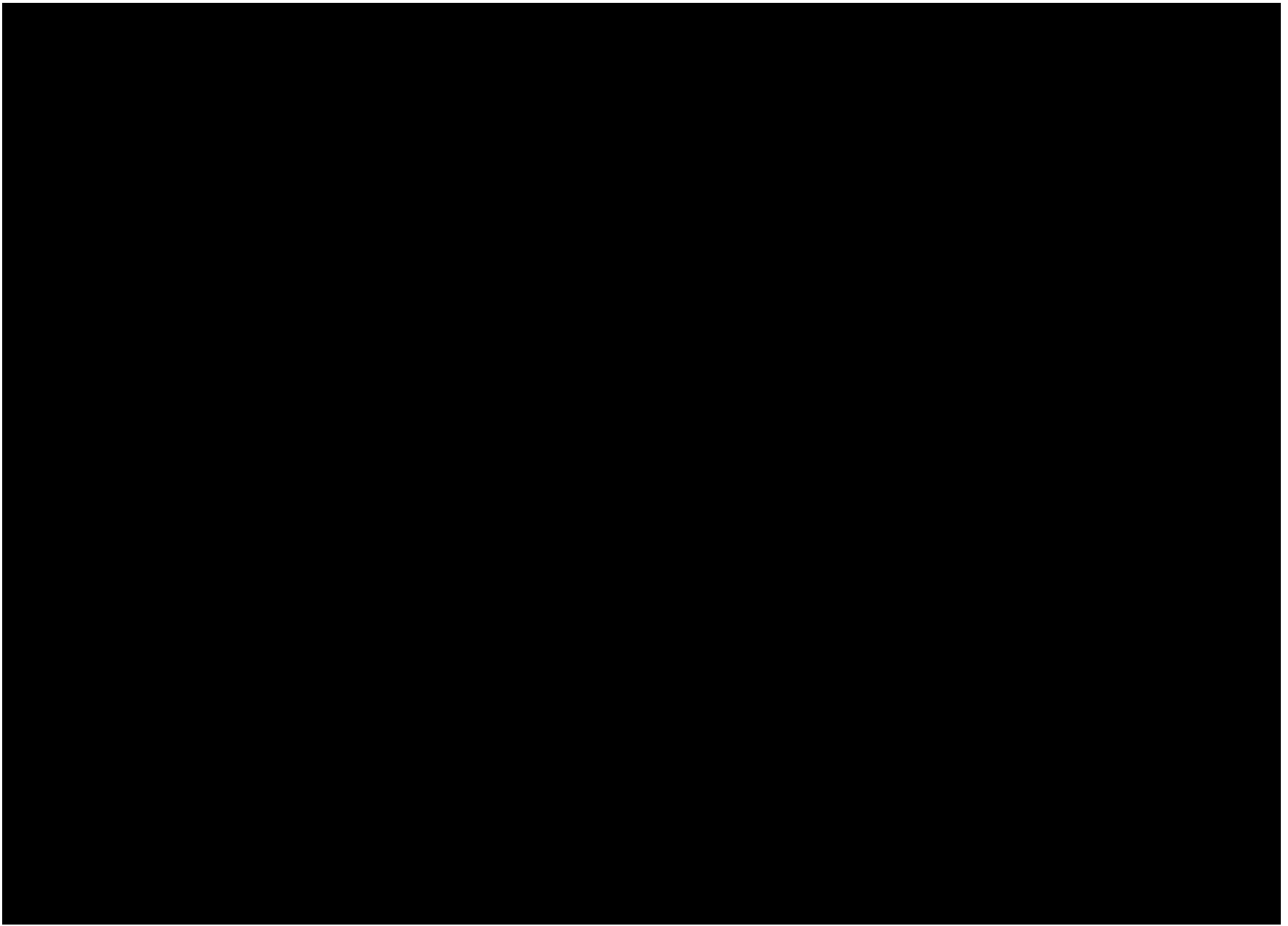
7) connectivity



Collocation (cont.)

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.

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



Practical Activities



Collocations in context with #LancsBox: Collocation graphs and networks

T **Task 1.** Create graphs. Work with the LOB corpus.

- a) Build a collocation graph (first-order collocates) around the word *time* using MI score and the default settings.



- b) How many collocates does the graph display? Are all of them useful?
- c) Change the default settings as indicated in the figure below (MI = 5 and above) and search for the node *time* again.



How many results did you get this time?

d) Which of the collocates occur predominantly to the left of the node *time* and which ones to the right?

Left:

.....

Right:

.....

e) Some of the collocates of *time* such as *t*, *kungo* might not be completely transparent. Use the right-click function to obtain concordances (KWIC pop-up) and explain these collocates.

t is used as

.....

kungo is used as

.....

Collocations in context with #LancsBox: Collocation graphs and networks

T **Task 1.** Create graphs. Work with the LOB corpus.

- a) Build a collocation graph (first-order collocates) around the word *time* using MI score and the default settings.



- b) How many collocates does the graph display? Are all of them useful?
315 collocates; no, this is an example of an overpopulated graph.

- c) Change the default settings as indicated in the figure below (MI = 5 and above) and search for the node *time* again.



How many results did you get this time? 35

d) Which of the collocates occur predominantly to the left of the node *time* and which ones to the right?

Left:

e.g. *at, long, first, short, same, spare, waste, twenty* etc.....

Right:

e.g. *ago, speak, saved, arrived, washing, erect, constants, etc.*.....

e) Some of the collocates of *time* such as *t, kungo* might not be completely transparent. Use the right-click function to obtain concordances (KWIC pop-up) and explain these collocates.

t is used as *a mathematical term (time t)*.....

kungo is used as *proper name in General fiction*.....