Stylistika IV

ZS 2024

Obsah

- kolokace

- literatura
  - McIntyre and Walker. 2019. Corpus Stylistics. (164nn)

Aktualizace a její projevy

- dominance vybraných lexémů

Aktualizace a její projevy

- dominance vybraných lexémů

- slova nestojí izolovaně

Aktualizace a její projevy

- dominance vybraných lexémů

- slova nestojí izolovaně ⟶ kolokace

Kolokace

- srov. KOLOKACE v NESČ
  - https://www.czechency.org/slovnik/KOLOKACE

- Wikipedia
  - https://en.wikipedia.org/wiki/Collocation

Kolokace

- srov. KOLOKACE v NESČ
  - https://www.czechency.org/slovnik/KOLOKACE

- ustálené
- neustálené

  - projev aktualizace?

Kolokace

- srov. KOLOKACE v NESČ
  - https://www.czechency.org/slovnik/KOLOKACE

- ustálené
- neustálené
  - projev aktualizace?
  - *ubíjet*

Kolokace

- srov. KOLOKACE v NESČ
  - https://www.czechency.org/slovnik/KOLOKACE

- ustálené
- neustálené

  - projev aktualizace?

  - *ubíjet*
    - *ubíjel mě stále novými kolokacemi*

Kolokace

- jak poznat „sílu“, resp. „(ne)ustálenost“ kolokací?

Kolokace

- jak poznat „sílu“, resp. „(ne)ustálenost“ kolokací?

- východiskem frekvence

Měření kolokability

- prostá frekvence
- porovnání s očekávanou náhodnou frekvencí
- asociační míry

Prostá frekvence

- lemma: *Ostrava*

- L1 kolokace

	lemma	f
1	v	2170
2	z	444
3	do	321
4	Baník	224
5	a	219
6	of	141
7	moravský	135
8	město	111
9	ArcelorMittal	83
10	centrum	74

R1 kolokace

	lemma	f
1	a	503
2	být	279
3	se	189
4	v	145
5	na	89
6	mít	64
7	i	52
8	do	51
9	s	50
10	nebo	34

Kolokace

- **prosté frekvence** v jistém smyslu **zavádějící**:

Kolokace

- **prosté frekvence** v jistém smyslu **zavádějící**:

- frekvence jevů ze stejného souboru (100 mil. korpus)

    - AB = 1000 výskytů
    - CD = 1000 výskytů
    - EF = 10 000 výskytů

Která kolokace je nejsilnější?

1. AB = 1000 výskytů
2. CD = 1000 výskytů
3. EF = 10 000 výskytů

- A_samost. = 10 000
- B_samost. = 10 000
- C_samost. = 1 000 000
- D_samost. = 1 000 000
- E_samost. = 1 000 000
- F_samost. = 1 000 000

Která kolokace je nejsilnější?

1. AB = 1000 výskytů
2. CD = 1000 výskytů
3. EF = 10 000 výskytů

A: 10 % ze všech jeho výskytů s B

- A_samost. = 10 000
- B_samost. = 10 000
- C_samost. = 1 000 000
- D_samost. = 1 000 000
- E_samost. = 1 000 000
- F_samost. = 1 000 000

Která kolokace je nejsilnější?

1. AB = 1000 výskytů
2. CD = 1000 výskytů
3. EF = 10 000 výskytů

A: 10 % ze všech jeho výskytů s B

C: 0.1 % ze všech jeho výskytů s D

- A_samost. = 10 000
- B_samost. = 10 000
- C_samost. = 1 000 000
- D_samost. = 1 000 000
- E_samost. = 1 000 000
- F_samost. = 1 000 000

Která kolokace je nejsilnější?

1. AB = 1000 výskytů
2. CD = 1000 výskytů
3. EF = 10 000 výskytů

A: 10 % ze všech jeho výskytů s B

C: 0.1 % ze všech jeho výskytů s D

E: 1 % ze všech jeho výskytů s F

- A_samost. = 10 000
- B_samost. = 10 000
- C_samost. = 1 000 000
- D_samost. = 1 000 000
- E_samost. = 1 000 000
- F_samost. = 1 000 000

Kolokace

- *Ostrava* = 7 372

- *v* = 2 298 086
- *Baník* = 2 105
- *a* = 3 130 431
- *ArcelorMittal* = 174

	lemma	f
1	v	2170
2	z	444
3	do	321
4	Baník	224
5	a	219
6	of	141
7	moravský	135
8	město	111
9	ArcelorMittal	83
10	centrum	74

Porovnání s očekávanou náhodnou frekvencí

- očekávané frekvence (f_o)

$$f_o = \frac{f_n \cdot f_k}{N}$$

f_n … frekvence daného slova

f_k … frekvence kolokátu

N … počet slov v korpusu

Porovnání s očekávanou náhodnou frekvencí

$$f_o = \frac{f_{Ostrava} \cdot f_v}{100000000} = \frac{7372 \cdot 2298086}{100000000} = 169{,}41$$

	lemma	f
1	v	2170
2	z	444
3	do	321
4	Baník	224
5	a	219
6	of	141
7	moravský	135
8	město	111
9	ArcelorMittal	83
10	centrum	74

Porovnání s očekávanou náhodnou frekvencí

$$f_o = \frac{f_{Ostrava} \cdot f_v}{100000000} = \frac{7372 \cdot 2298086}{100000000} = 169{,}41$$

$$f_o = \frac{f_{Ostrava} \cdot f_{Banik}}{100000000} = \frac{7372 \cdot 2105}{100000000} = 0{,}16$$

	lemma	f
1	v	2170
2	z	444
3	do	321
4	Baník	224
5	a	219
6	of	141
7	moravský	135
8	město	111
9	ArcelorMittal	83
10	centrum	74

Porovnání s očekávanou náhodnou frekvencí

$$f_o = \frac{f_{Ostrava} \cdot f_v}{100000000} = \frac{7372 \cdot 2298086}{100000000} = 169{,}41$$

$$f_o = \frac{f_{Ostrava} \cdot f_{Banik}}{100000000} = \frac{7372 \cdot 2105}{100000000} = 0{,}16$$

$$f_o = \frac{f_{Ostrava} \cdot f_a}{100000000} = \frac{7372 \cdot 3130431}{100000000} = 230{,}78$$

	lemma	f
1	v	2170
2	z	444
3	do	321
4	Baník	224
5	a	219
6	of	141
7	moravský	135
8	město	111
9	ArcelorMittal	83
10	centrum	74

Porovnání s očekávanou náhodnou frekvencí

$$f_o = \frac{f_{Ostrava} \cdot f_v}{100000000} = \frac{7372 \cdot 2298086}{100000000} = 169{,}41$$

$$f_o = \frac{f_{Ostrava} \cdot f_{Banik}}{100000000} = \frac{7372 \cdot 2105}{100000000} = 0{,}16$$

$$f_o = \frac{f_{Ostrava} \cdot f_a}{100000000} = \frac{7372 \cdot 3130431}{100000000} = 230{,}78$$

$$f_o = \frac{f_{Ostrava} \cdot f_{ArcelorMittal}}{100000000} = \frac{7372 \cdot 174}{100000000} = 0{,}01$$

	lemma	f
1	v	2170
2	z	444
3	do	321
4	Baník	224
5	a	219
6	of	141
7	moravský	135
8	město	111
9	ArcelorMittal	83
10	centrum	74

Kolokace

- asociační míry
  - NESČ
    - https://www.czechency.org/slovnik/ASOCIA%C4%8CN%C3%8D%20M%C3%8DRA
  - Český národní korpus
    - https://wiki.korpus.cz/doku.php/pojmy:asociacni_miry

Kolokace

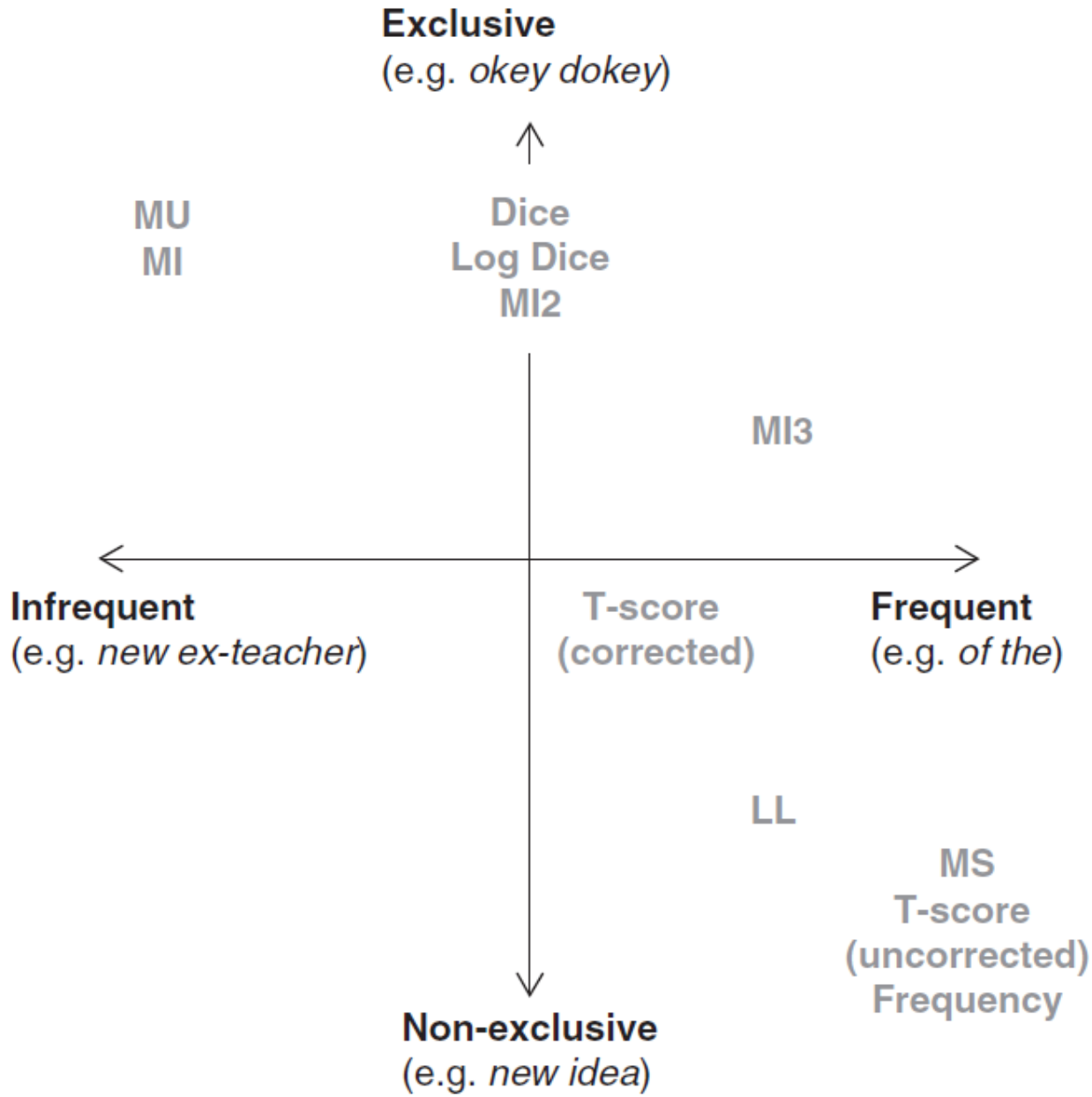
- asociační míry
  - matematické postupy používané pro vyhledání kolokací v korpusu
  - většinou vycházejí ze statistického testování hypotéz

Kolokace

- asociační míry
  - matematické postupy používané pro vyhledání kolokací v korpusu
  - většinou vycházejí ze statistického testování hypotéz
- záporné hodnoty asociačních měr vykazují negativní asociace, tj. vzájemné „odpuzování“

Kolokace

- asociační míry
  - matematické postupy používané pro vyhledání kolokací v korpusu
  - většinou vycházejí ze statistického testování hypotéz
- záporné hodnoty asociačních měr vykazují negativní asociace, tj. vzájemné „odpuzování“
- různé způsoby měření
  - MI-score
  - T-score
  - MIt-score
  - Dice & logDice
  - Log likelihood
  - Min. sensitivity
  - Chi-kvadrát
  - z-score


Exclusive
(e.g. okey dokey)
MU
MI
Dice
Log Dice
MI2
MI3
Infrequent
(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)

Výpočet asociačních měr

- f(x), f(y) pro frekvenci slov x a y

Výpočet asociačních měr

- f(x), f(y) pro frekvenci slov x a y
- f(xy) pro frekvenci spojení xy

Výpočet asociačních měr

- f(x), f(y) pro frekvenci slov x a y
- f(xy) pro frekvenci spojení xy
- N pro velikost korpusu (počet tokenů)

Výpočet asociačních měr

- f(x), f(y) pro frekvenci slov x a y
- f(xy) pro frekvenci spojení xy
- N pro velikost korpusu (počet tokenů)

N = 100 000 000

Výpočet asociačních měr

- f(x), f(y) pro frekvenci slov x a y
- f(xy) pro frekvenci spojení xy
- N pro velikost korpusu (počet tokenů)

N = 100 000 000

f(dobrý) = 116 272

Výpočet asociačních měr

- f(x), f(y) pro frekvenci slov x a y
- f(xy) pro frekvenci spojení xy
- N pro velikost korpusu (počet tokenů)

N = 100 000 000

f(dobrý) = 116 272
f(den) = 111 733

Výpočet asociačních měr

- f(x), f(y) pro frekvenci slov x a y
- f(xy) pro frekvenci spojení xy
- N pro velikost korpusu (počet tokenů)

N = 100 000 000

f(dobrý) = 116 272
f(den) = 111 733
f(dobrý den) = 1 626

Výpočet asociačních měr

- f(x), f(y) pro frekvenci slov x a y
- f(xy) pro frekvenci spojení xy
- N pro velikost korpusu (počet tokenů)

N = 100 000 000

f(dobrý) = 116 272
f(den) = 111 733
f(dobrý den) = 1 626

f(dobrý) = 116 272

Výpočet asociačních měr

- f(x), f(y) pro frekvenci slov x a y
- f(xy) pro frekvenci spojení xy
- N pro velikost korpusu (počet tokenů)

N = 100 000 000

f(dobrý) = 116 272
f(den) = 111 733
f(dobrý den) = 1 626

f(dobrý) = 116 272
f(jitro) = 446

Výpočet asociačních měr

- f(x), f(y) pro frekvenci slov x a y
- f(xy) pro frekvenci spojení xy
- N pro velikost korpusu (počet tokenů)

N = 100 000 000

f(dobrý) = 116 272
f(den) = 111 733
f(dobrý den) = 1 626

f(dobrý) = 116 272
f(jitro) = 446
f(dobrý jitro) = 111

MI-score

$$MI = log_2 \frac{p(xy)}{p(x)p(y)} = log_2 \frac{Nf(xy)}{f(x)f(y)}$$

$$MI_{dobrý\ den} = log_2 \frac{100\ 000\ 000 \cdot 1626}{116\ 272 \cdot 111\ 733} = 3{,}65$$

$$MI_{dobré\ jitro} = log_2 \frac{100\ 000\ 000 \cdot 111}{116\ 272 \cdot 446} = 7{,}74$$

T-score

$$T(xy) = \frac{f(xy) - \frac{f(x)f(y)}{N}}{\sqrt{f(xy)}}$$

$$T(dobrý\ den) = \frac{1626 - \frac{116\,272 \cdot 111\,733}{100\,000\,000}}{\sqrt{1626}} = 37{,}1$$

$$T(dobré\ jitro) = \frac{111 - \frac{116\,272 \cdot 446}{100\,000\,000}}{\sqrt{111}} = 10{,}49$$

Kolokační grafy

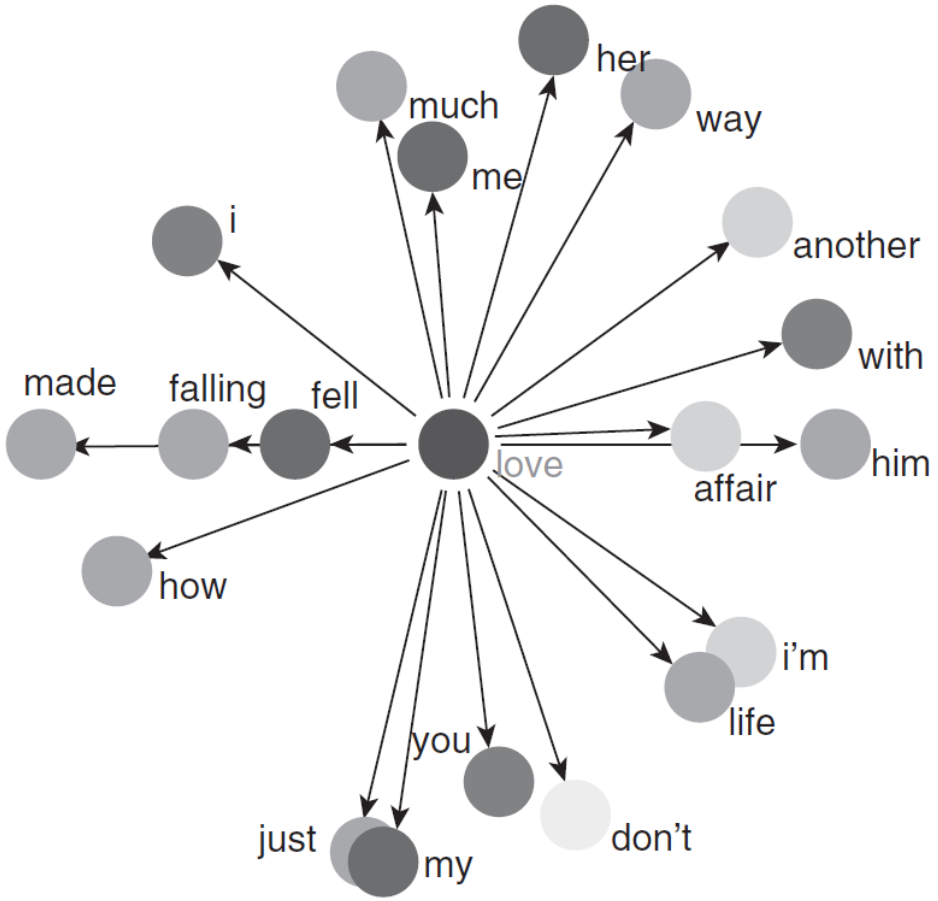


Figure 3.2 *Collocation graph: 'love' in BE06 (10a – log Dice (7), L3–R3, C5–NC5)*

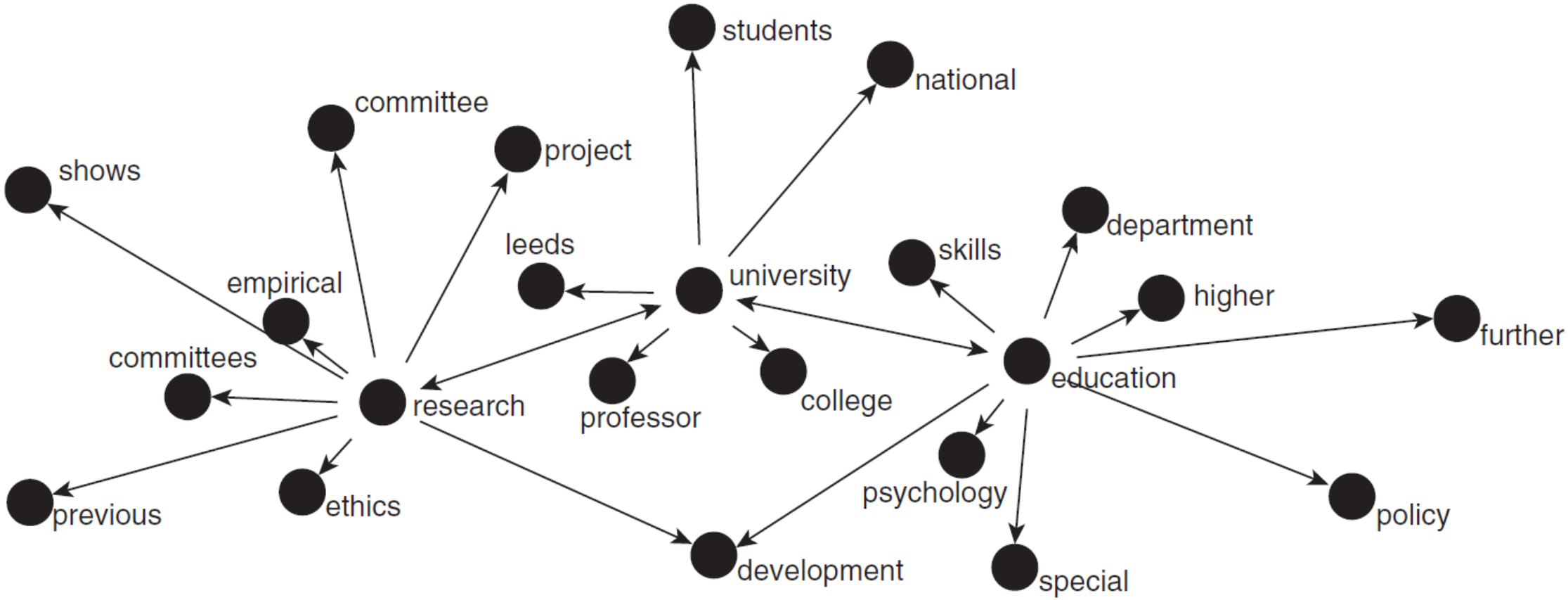


Figure 3.5 *Collocation network of 'university' based on BE06 (3b–MI(3), L5–R5, C8–NC8)*