Education as an over-represented topic in the ICE corpora?

John Newman & Georgie Columbus
University of Alberta

15th Conference of the International Association for World Englishes
Cebu City, Philippines, 22 to 24 October, 2009

our talk

- The ICE corpora
- Lexical items vs. Lexical sets
- Education domain in ICE corpora
- Syntax-lexis interface
- Conclusion

the ICE corpora

- Initiated by Sidney Greenbaum in 1989 for study of English varieties
- 1 million words in each corpus (60% spoken)
- 8 corpora finished: IND, GB, IRE, PHIL, SIN, HK, JAM, NZ; and one nearing completion: CAN

can ICE inform on lexis?

- "...ICE-GB was designed primarily as a resource for syntactic studies, not for lexical studies."

  Nelson, Wallis, and Aarts (2002)

- "A 200,000-word subcorpus is adequate for most studies of grammar and some studies of lexis, but is insufficient, for example, for lexical investigations involving low frequency words."

  Granger (1996)

methodology

- Used spoken face-to-face subcorpus from the 9 corpora (180,000 words in each subcorpus)
- Devised a comparably-sized set of lexical items for each domain
    - Sports, education, legal, business/finance, work, health, government, climate and arts
- Lists adapted from vocabulary lists for ESL (plus collocates of these items)

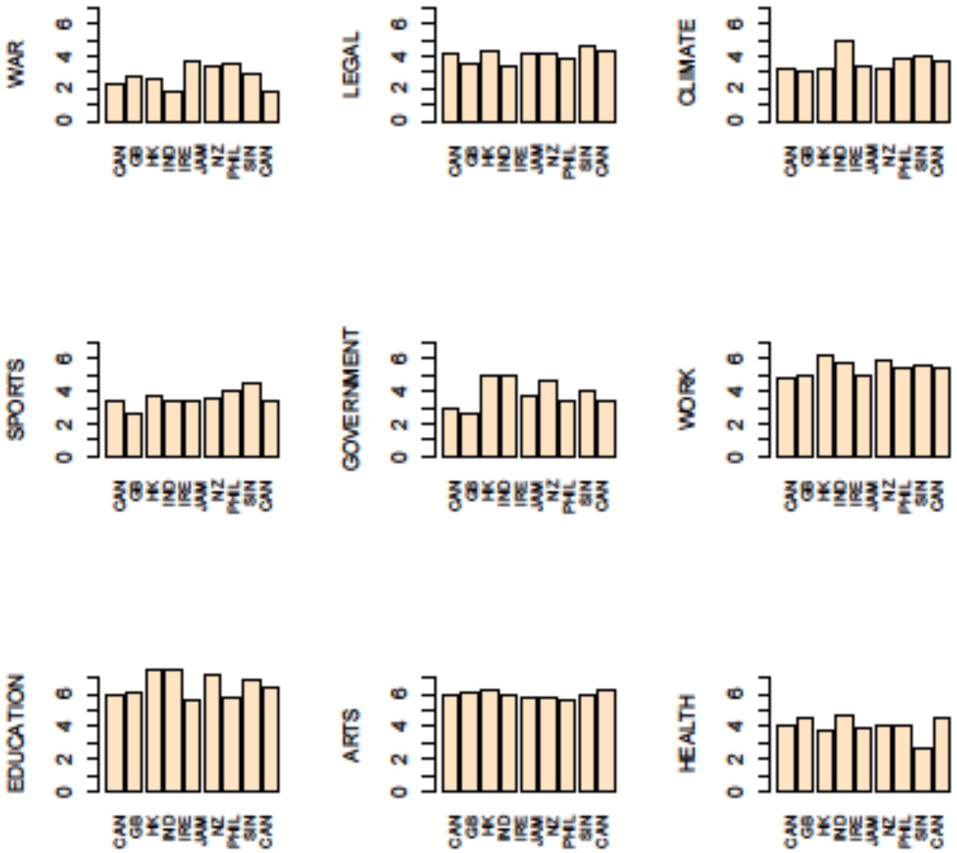
lexical set for the WORK domain

administrator, application, apply, apprentice, assistant, benefits, blue, collar, boss, business, career, CEO, colleague, convention, coworker, customer, service, demotion, discrimination, downsizing, experience, factory, fax, fired, fires, firing, firm, hire, hirers, hiring, human, resources, inventory, job, job, offer, labour, laid, off, layoff, layoffs, management, manager, networking, nightshift, paycheck, paycheque, pays, payslip, president, production, promotion, public, relations, qualified, receptionist, secretary, shift, shiftwork, staff, supervisor, switchboard, synergy, trade, training, white, collar, work, experience

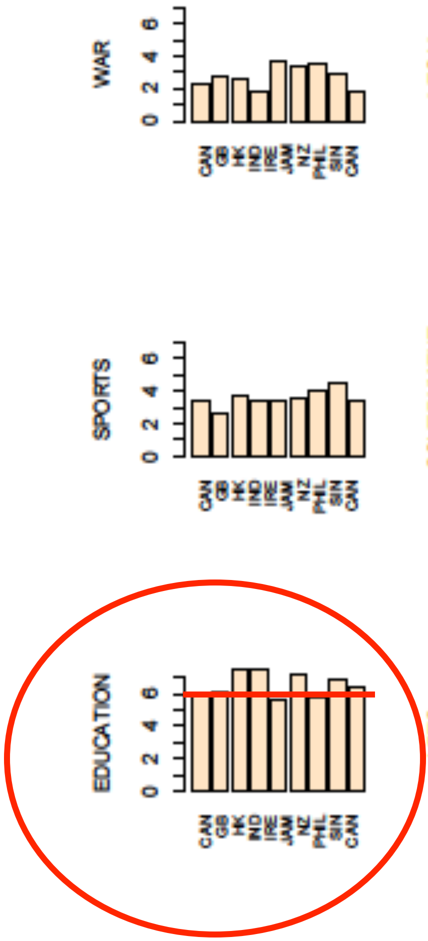
size of lexical sets

domain	no. of seed words
arts	50
educ	66
business/finance	74
climate	50
government	50
health	64
legal	53
sports	57
war	50
work	64

frequency of domains in 9 corpora



frequency of domains in 9 corpora

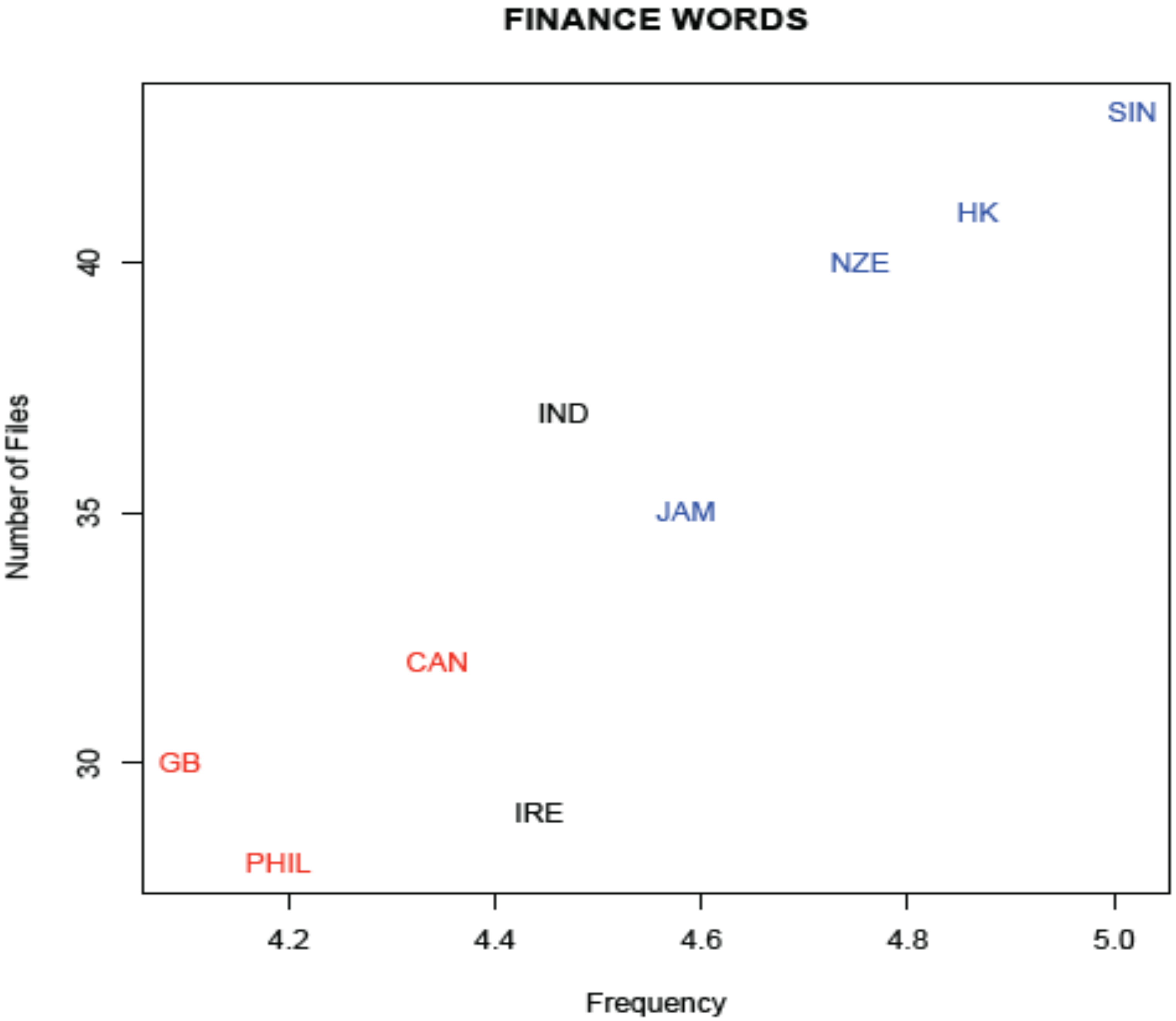


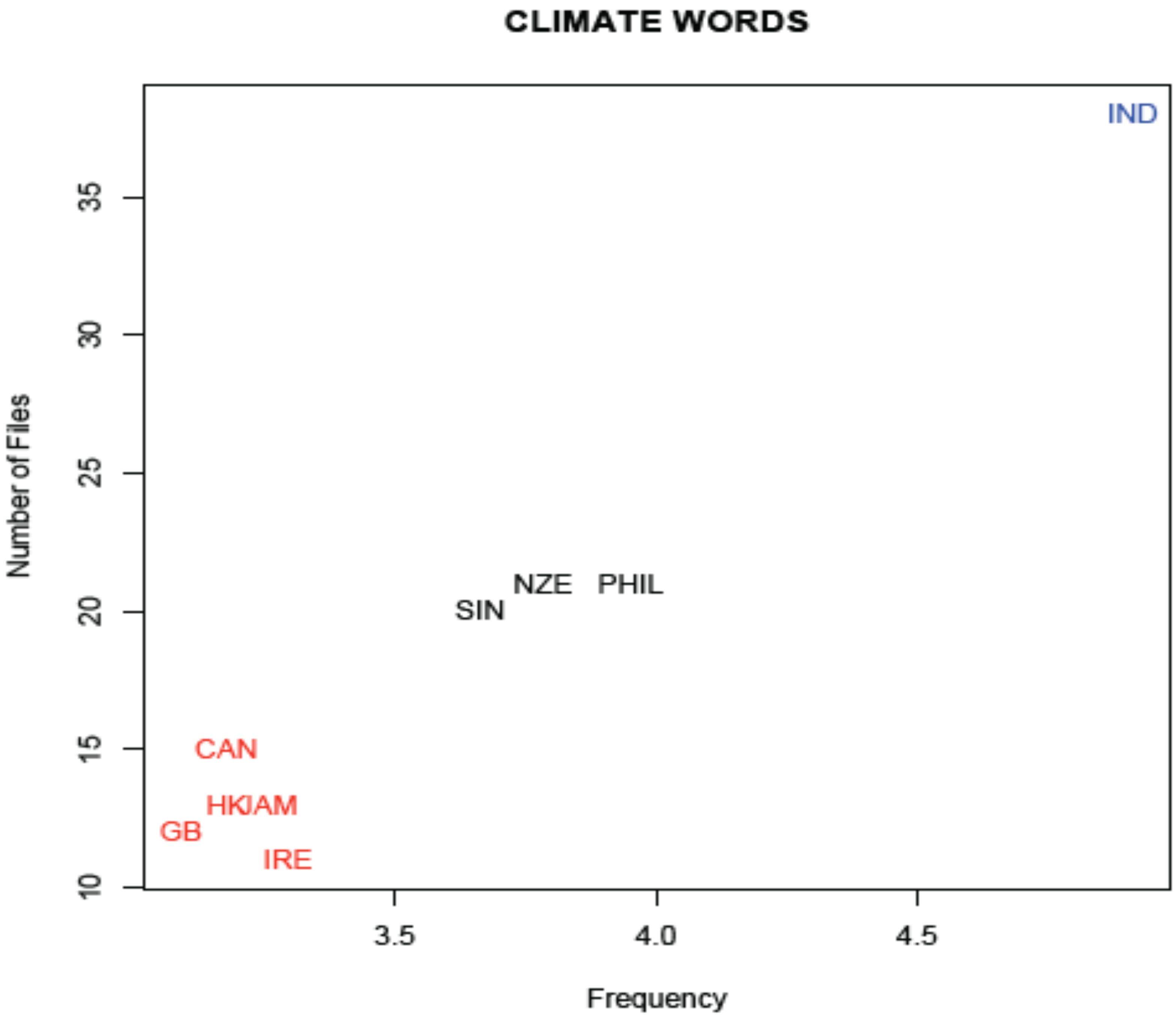
6 = log (403)
5 = log (148)

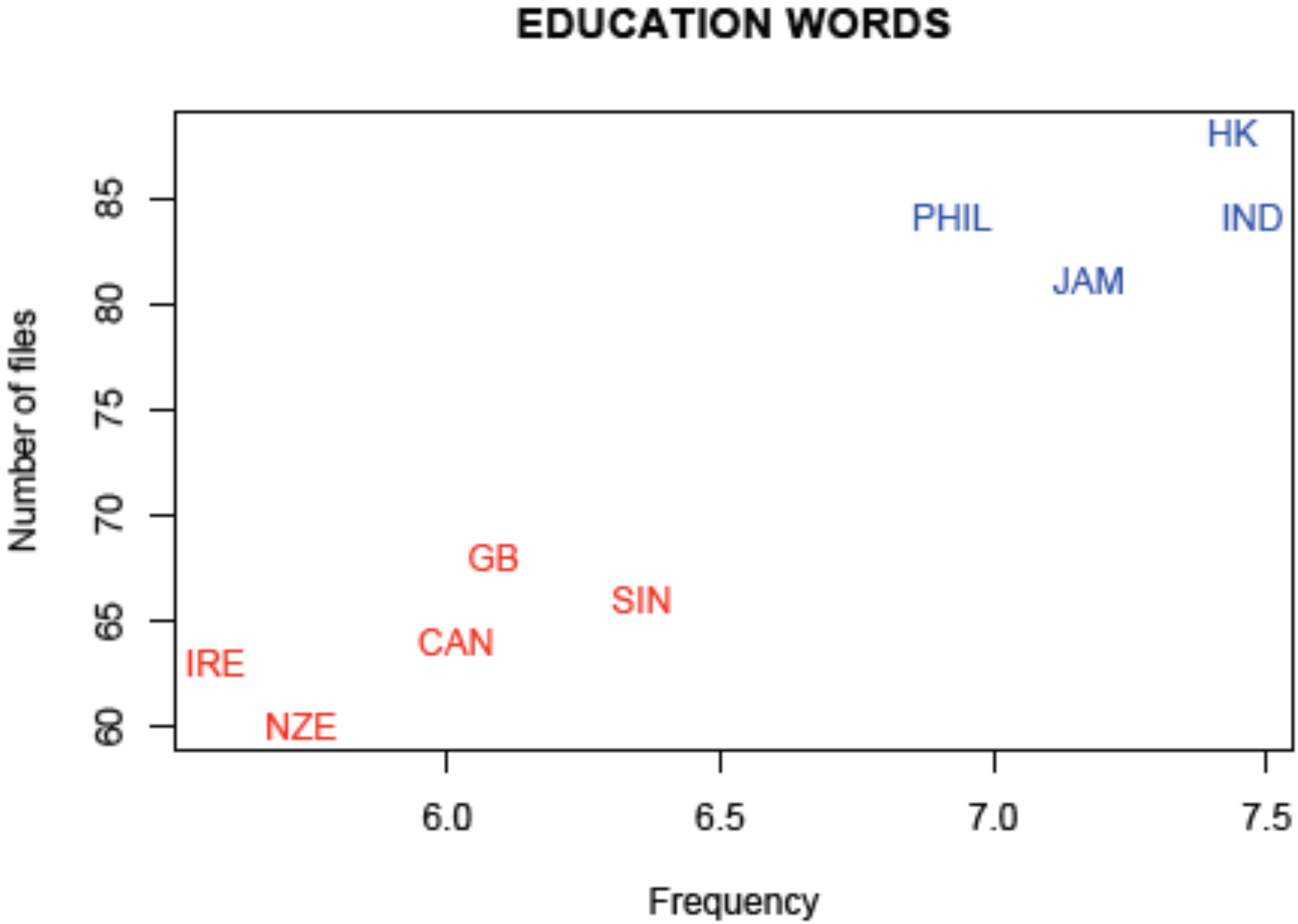
frequency and dispersion

- frequency of a topic in each ICE corpus
  - Aggregate the individual frequencies of all words in a domain across all files (each occurrence was checked manually)

- dispersion of topic throughout each ICE corpus
  - Simply count the number of files in which any domain word appears


FINANCE WORDS
SIN
HK
NZE
IND
JAM
CAN
GB
IRE
PHIL
40
35
30
Number of Files
4.2
4.4
4.6
4.8
5.0
Frequency


CLIMATE WORDS
IND
NZE
PHIL
SIN
CAN
HKJAM
GB
IRE
35
30
25
20
15
10
Number of Files
3.5
4.0
4.5
Frequency


EDUCATION WORDS
Number of files
60
65
70
75
80
85
6.0
6.5
7.0
7.5
Frequency
HK
PHIL
IND
JAM
GB
SIN
CAN
IRE
NZE

focus on EDUCATION Domain

- Relatively high frequency of occurrence
- Relatively even distribution
- IND, HK, JAM, and PHIL: higher frequency and use in larger no. of files
- GB, SIN, CAN, IRE, NZE: lower frequency and use in smaller no. of files

an aside: educational bias and data collection?

1. Overreliance on the education profession (teachers, students, lecturers, research or teaching assistants etc.)?
   - ICE-NZ: 28.6% of speakers in this category
   - ICE-Can 43.7% of speakers in this category

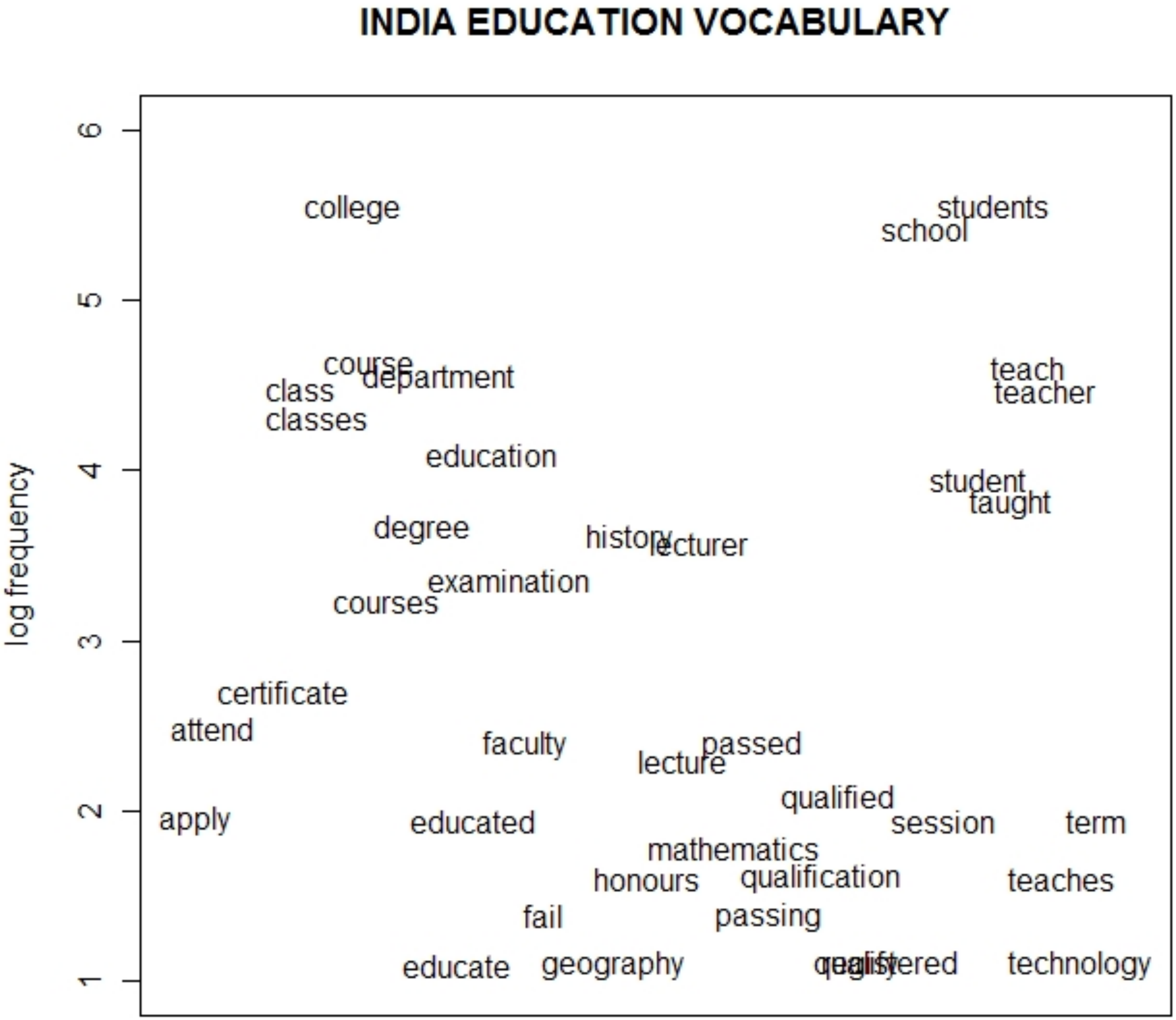
2. Education as a starter topic for conversations?
   - ICE-IND and ICE-JA have education or educational occupation as a starter topic in many files

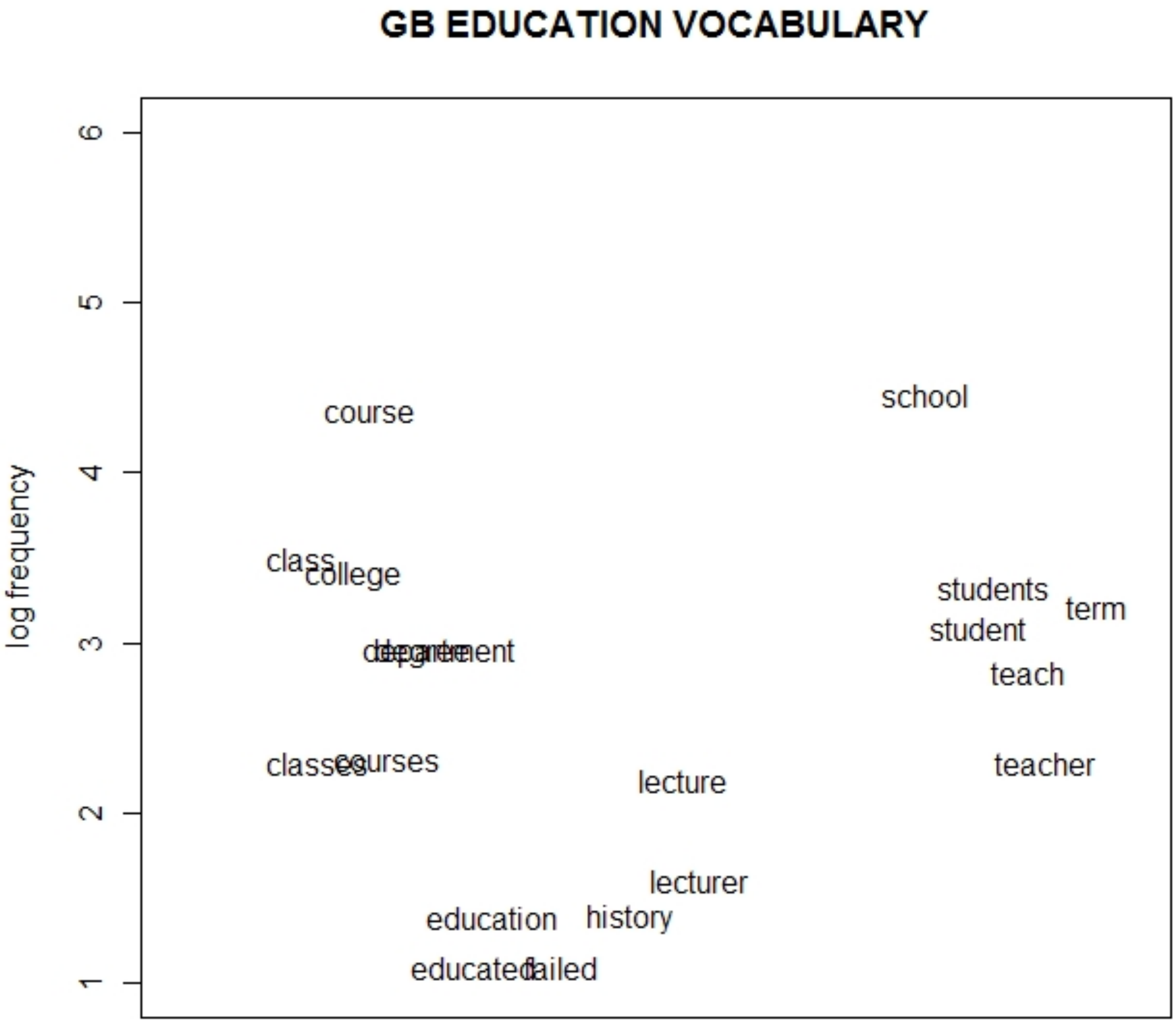
an aside: beginning of an ICE-INDIA conversation

A: So when did you finish your M A degree

B: In fact I finished my M A some uh from the University of Jabalpur Madhya Pradesh

A: Uhm I see
That means uh you finished it in the university itself
Which year could you tell me the year


INDIA EDUCATION VOCABULARY
log frequency
6
5
4
3
2
1
college
students
school
course
department
class
classes
teach
teacher
education
student
taught
degree
history
lecturer
examination
courses
certificate
attend
faculty
passed
lecture
qualified
apply
educated
session
term
mathematics
honours
qualification
teaches
fail
passing
educate
geography
qualify
registered
technology


GB EDUCATION VOCABULARY
log frequency
6
5
4
3
2
1
course
school
class
college
students
term
student
department
degree
teach
classes
courses
teacher
lecture
lecturer
education
history
educated
failed

EDUCATION core words (in all ICE corpora)

class, classes, college, course, courses degree, department, educated, education, lecture, school, student, students, teach, teacher, term

reference corpora for comparison

- BNC Baby v.1 conversation
  - 1 million words

- BNC World Edition, spoken demographic
  - 4.2 million words

ICE vs. BNC Baby
aggregated core EDUCATION words

	ICE 180,000 words	BNC Baby 1 million words	LL	Overuse in ICE?
IND	1,493	717	3,067	yes
HK	1,389	717	2,760	yes
JAM	1,124	717	2,003	yes
PHIL	898	717	1,396	yes
SIN	481	717	432	yes
GB	404	717	291	yes
CAN	363	717	223	yes
NZ	280	717	106	yes
IRE	234	717	56	yes

LL score of 3.8 or higher is significant at $p < 0.05$; LL score of 6.6 or higher is significant at $p < 0.01$.
Paul Rayson's Log Likelihood Calculator: http://ucrel.lancs.ac.uk/llwizard.html

ICE vs. BNC spok. dem.
aggregated core EDUCATION words

	ICE 180,000 words	BNC spok dem 4.2 million words	LL	Overuse in ICE?
IND	1493	3,674	3,630	yes
HK	1389	3,674	3,229	yes
JAM	1124	3,674	2,262	yes
PHIL	898	3,674	1,513	yes
SIN	481	3,674	402	yes
GB	404	3,674	253	yes
CAN	363	3,674	185	yes
NZE	280	3,674	74	yes
IRE	234	3,674	31	yes

ICE-INDIA vs. BNC spoken demographic individual core EDUCATION words

	ICE -IND 180,000 words	BNC spok dem 4.2 million words	LL	Overuse in ICE-IND?
students	261	71	1328	yes
class	89	258	716	yes
department	94	13	522	yes
college	256	284	516	yes
teach	100	88	386	yes
course	102	165	367	yes
teacher	88	260	190	yes
student	52	66	176	yes
education	60	112	170	yes
school	228	2022	150	yes
courses	25	36	76	yes
classes	75	46	53	yes
degree	39	32	47	yes
educated	7	12	21	yes
lecture	10	41	17	yes
term	7	168	0	ns

Lexis-Syntax

- What are the implications of an overuse of EDUCATION lexis for (lexically sensitive) syntactic study?

- We'll explore <to NP> constructions

Two simple percentages

$$\text{Attraction} = \frac{\text{frequency of X in a pattern}}{\text{frequency of pattern}} \times 100$$

$$\text{Reliance} = \frac{\text{frequency of X in a pattern}}{\text{frequency of X in corpus}} \times 100$$

Hans-Jörg Schmid. (in press). Does frequency in text really instantiate entrenchment in the cognitive system? And do we have a quantitative grip on either of them? In Dylan Glynn and Kersin Fischer (eds.), *Quantitative Methods in Cognitive Semantics*. Berlin/New York: Mouton de Gruyter.

Attraction of *school* to <to X N>

	Freq of <to X school>	Freq of <to_PREP>	Attraction
ICE-IND direct conv.	23	1012	2.27
ICE-GB direct conv.	22	1059	2.08
BNCBaby spok dem	88	5193	1.69
BNC spok dem	374	23449	1.59

ICE-INDIA compared with BNCBaby: chi-squared = 1.2435, df = 1, p-value = 0.2648, ns

Reliance of *school* on <to X *school*>

	Freq of <to X *school*>	Freq of *school*	Reliance
ICE-IND direct conv.	23	226	10.18%
ICE-GB direct conv.	22	88	25.00%
BNCBaby spok dem	88	415	21.20%
BNC spok dem	374	2022	18.50%

ICE-INDIA compared with BNCBaby : chi-squared = 8.383, df = 1, $p<0.005$, significant

Reliance of *school* on <to X *school*>

	Freq of	Freq of	
	<to X *school*>	*school*	Reliance
ICE-IND direct conv.	23	226	10.18%
ICE-GB direct conv.	22	88	25.00%
BNCBaby spok dem	88	415	21.20%
BNC spok dem	374	2022	18.50%

ICE-INDIA compared with BNCBaby : chi-squared = 8.383, df = 1, p<0.005, significant

Collostructional Analysis

- word.freq: frequency of the word in the corpus
- obs.freq: observed frequency of the word with/in TO
- exp.freq: expected frequency of the word with/in TO
- faith: percentage of how many instances of the word occur with/in TO
- relation: relation of the word to TO [requires also total frequency of PREP constructions in the corpus]
- coll.strength: index of collocational/collostructional strength:
  -log(Fisher exact, 10), the higher, the stronger

Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction between words and constructions. International Journal of Corpus Linguistics 8.2:209-43

Gries, Stefan Th. 2004. Coll.analysis 3. A program for R for Windows 2.x.

collostructional analysis: <to NP> in BNC Baby spoken dem.

words	word.freq	obs.freq	exp.freq	faith	relation	coll.strength
school	415	88	47.69	0.212	attraction	8
office	128	16	14.71	0.125	attraction	0.4
education	15	2	1.72	0.1333	attraction	0.3
people	996	23	114.47	0.0231	repulsion	27.1
fact	176	0	20.23	0	repulsion	9.4
children	190	5	21.84	0.0263	repulsion	5.1
teacher	54	1	6.21	0.0185	repulsion	2
city	30	1	3.45	0.0333	repulsion	0.9
students	16	0	1.84	0	repulsion	0.8
college	36	2	4.14	0.0556	repulsion	0.7

Coll.strength>3 => p<0.001; coll.strength>2 => p<0.01; coll.strength>1.30103 => p<0.05.

collostructional analysis:
<to NP> in ICE-GB direct conv.

words	word.freq	obs.freq	exp.freq	faith	relation	coll.strength
school	88	22	8.67	0.25	attraction	4.5
teacher	10	2	0.98	0.2	attraction	0.6
college	30	4	2.95	0.1333	attraction	0.5
people	448	17	44.12	0.0379	repulsion	6.1
fact	101	0	9.95	0	repulsion	4.6
students	29	0	2.86	0	repulsion	1.3
children	36	1	3.55	0.0278	repulsion	0.9
city	11	0	1.08	0	repulsion	0.5
education	4	0	0.39	0	repulsion	0.2
office	13	1	1.28	0.0769	repulsion	0.2

Coll.strength>3 => p<0.001; coll.strength>2 => p<0.01; coll.strength>1.30103 => p<0.05.

Collostructional Analysis: <to NP> in ICE-INDIA direct conv.

words	word.freq	obs.freq	exp.freq	faith	relation	coll.strength
school	226	23	16.55	0.1018	attraction	1.2
office	39	6	2.86	0.1538	attraction	1.2
city	74	7	5.42	0.0946	attraction	0.5
students	262	20	19.19	0.0763	attraction	0.3
people	554	16	40.57	0.0289	repulsion	5.4
fact	90	0	6.59	0	repulsion	3
teacher	88	2	6.44	0.0227	repulsion	1.4
education	57	1	4.17	0.0175	repulsion	1.1
college	257	14	18.82	0.0545	repulsion	0.8
children	87	6	6.37	0.069	repulsion	0.3

Coll.strength>3 => p<0.001; coll.strength>2 => p<0.01; coll.strength>1.30103 => p<0.05.

Uses of *school* in ICE-INDIA

1. and by the time I reach to school uhm you know what I'll be and
2. I'll get my confidence then I'll go School
3. So when will you be joining school
4. They are having the plus two plus plus system of Indian school we call it as
5. Whenever I enter into school I'll be very cheerful with the children

conclusion

- By extending a lexical study to sets of words in a domain, even small corpora such as the ICE corpora can inform on topic/content preferences in corpora

- Overuse in the lexis of a domain in a corpus does not imply overuse of that lexis in every construction type

references

Granger, Sylviane. (1996). Learner English around the world. In Sidney Greenbaum (ed.), *Comparing English Worlwide: The International Corpus of English,* pp. 13- 24. Oxford: Clarendon Press.

Nelson, Gerald, Sean Wallis, and Bas Aarts. (2002). *Exploring natural language: Working with the British Component of the International Corpus of English.* Amsterdam and Philadelphia: John Benjamins.