University of Zurich^UZH

The Detection of Learner Difficulties from Unannotated Corpora

UZH-GE Workshop on Computers in L2 Learning & Assessment

PD Dr. Gerold Schneider

Computational Linguistics & English Department, University of Zurich
Lehrbeauftragter an der TU Dortmund

Presentation mainly based on

Schneider, Gerold and Gaëtanelle Gilquin. (2016). “Detecting Innovations in a Parsed Corpus of Learner English”. *International Journal of Learner Corpus Research*. 2(2). ISSN 2215-1478.

Schneider, Gerold and Johannes Graën (2018). “NLP Corpus Observatory – Looking for Constellations in Parallel Corpora to Improve Learners’ Collocational Skills.“ Proceedings of NLP4CALL, Stockholm, November 7.

The Detection of Learner Difficulties from Unannotated Corpora

They have to cope with life's problems and difficulties, and to realize the reasons why they decided to get *involved into* crimes. **(ICLE ITTO 1019)**

Contents

1. Introduction
2. Collocations
3. Helping Learners: Non-literal Translations show Areas of Difficulty
4. Translational Scapes
5. Characteristics of L2

With many thanks to

Johannes Graën, Gaëtanelle Gilquin, Gintare Grigonyte, Hans Martin Lehmann

1. Introduction

1.1. Errors & Non-native-like features of EFL

Non-native-like features in EFL production are interesting

- Cognitive challenges → cognitive linguistics
- Learner difficulties → help learners

EFL features are not only errors:

- More than typos, but lexico-grammatical patterns
- They are used repeatedly, partly reach collocational status
- Can be due to L1 transfer or cognitive/semantic analogy
- Language is an inherently gradient system

The application of technologies like parsing to learner corpora helps automate the detection of non-native-like features (data-driven methods)

1.2 Verb+PP combinations

- Ng et al. (2014): 3 most frequent error types by learners of English:
  - Wrong collocation or idiom: 14.2 to 14.4%
  - Article error: 13.3 to 13.9%
  - Preposition Error: 8.8 to 11.7%
- Prepositions exhibit a high rate of innovation, both in ESL and EFL.
  - ESL: e.g. Indian English, presents a high degree of innovation in its use of prepositional verbs (Mukherjee & Hoffmann 2006)
  - EFL: Prepositions are difficult to acquire for non-native speakers, (see Gilquin & Granger 2011: 59-60).
- Routinisation is particularly difficult for learners: “A focus of the lexical approach to language pedagogy is teaching collocations .. Such knowledge is evidently more important than individual words themselves” (McEnery & Xiao 2001:368)
- Understandable VS native-like English (Pawley & Syder 1983)

1.2 Verb+PP: including adjectives & phrasal verbs

- Phrasal verbs represent “one of the most notoriously challenging aspects of English language instruction” (Gardner & Davies 2007: 339)
- Often the new combination involves a confusion between the two: e.g. *depend on* vs*. depend from*
- We also include adjective + PP combinations, as they, too, have collocational status. For example, Benson et al. (2009) recognise *adjective* + *preposition* as an independent category in addition to *verb* + *preposition* (and *noun* + *preposition*, e.g. in nominalisations, which we have not included)*. Adjective* + *preposition* combinations are often similarly difficult to acquire for learners of English.

1.3 Syntactic Parsing

Parsing technology has now matured enough to deliver syntactically annotated large corpora with error rates that are acceptably low for these types of research (van Noord and Bouma 2009). We have parsed the BNC and other large corpora using a dependency parser (Schneider 2008)

- Advantage of (semi-) automatic, parse-based methods: fast and corpus-driven, which may increase recall
- Disadvantage: error-rates are high, possibly higher in L2, which affects precision and recall. The small ICLE corpus
  - poses particular challenges to automated detection of rare collocations (recall),
  - while manual filtering of lists of suggested candidates is easily possible (precision).

1.4 Materials: the Corpora

- **EFL**: International Corpus of Learner English (ICLE; Granger et al. 2009). Corpus of learner English from university students with 16 different mother tongues. It contains 3.7 million words from essays of higher intermediate to advanced learners of English.
- **ENL**: written part of the British National Corpus (BNC; Aston & Burnard 1998). It contains 90 million words of written texts from a wide range of registers. We use it as a reference corpus of native British English.
- **Student Essays in ENL**: genre-matched corpus, compiled by the ICLE team: LOCNESS corpus. 320.000 words
- **Parallel Corpus**: EuroParl (Corrected & Structured Europarl Corpus; Graën, Batinic, and Volk (2014))

1.4 Materials: the Corpora

Linguistic Backgrounds in ICLE

```
$ wc -w *
  201265 BG_ALL.txt  %% Bulgarian
  493347 CN_ALL.txt  %% Chinese
  202651 CZ_ALL.txt  %% Czech
   96496 DB_ALL.txt  %% Dutch Belgian
  138863 DN_ALL.txt  %% Dutch Netherlands
  275610 FI_ALL.txt  %% Finnish
  227764 FR_ALL.txt  %% French
  231037 GE_ALL.txt  %% German
  224937 IT_ALL.txt  %% Italian
  198540 JP_ALL.txt  %% Japanese
  212205 NO_ALL.txt  %% Norwegian
  234620 PO_ALL.txt  %% Polish
  230385 RU_ALL.txt  %% Russian
  198486 SP_ALL.txt  %% Spanish
  200734 SW_ALL.txt  %% Swedish
  199840 TR_ALL.txt  %% Turkish
  199939 TS_ALL.txt  %% Tswana, South Africa
 3766719 total
```

1.5 Research Questions

Some Learner Corpora are error-tagged, but most are not.

Can we use them to detect errors?

1) Can the patterns of overuse which we observe with collocation statistics deliver combinations that are specific to EFL / ESL?
2) Does the method give us the tools to find more patterns than have been previously described?
3) Can we use parallel corpora to help us further?
4) Can we observe further characteristics of learner language?

2. Collocations

Some Learner Corpora are error-tagged, but most are not.

Can we use them to detect errors?

We want to detect general patterns, and particularly verb-PP combinations which

1) are frequent enough to reach collocation status

2) are collocations in L2

3) but not, or much less so, in L1

If we apply traditional collocation measures we fail to see 3)

Let’s first repeat collocations:

2.1 Collocation measures

[A collocation is defined as] a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components. (Choueka 1988)

Some criteria:

- Non-compositionality
  - meaning not compositional (e.g. “kick the bucket”)
- Non-substitutability
  - near synonyms cannot be used (e.g. “yellow wine”?)
- Non-modifiability
  - “kick the bucket”, “*kick the buckets”, “*kick the blue bucket”
- Non-literal translations
  - “red wine” <-> “vino tinto”, “take decisions” <-> “Entscheidungen treffen”
- Frequently occurring together, “mutually attracting each other”
  - easy to calculate, works surprisingly well

2.1 O/E (Observed divided by Expected, O over E)

- Probability that collocation (x,y) is due to chance [Expectation, independent events]: P(x) * P(y)
- P(x) = f(x)/N ; P(y) = f(y)/N [N = corpus size in words]
- Actual measurement [Observed]: P(x,y)
  P(x,y) = f(x,y) / N
- If the collocations is due to chance (independent) we expect
  P(x,y) =~ P(x)*P(y)
- If P(x,y)>>P(x)*P(y) then strong collocation
- If P(x,y)<<P(x)*P(y) then 'negative' collocation
- MI originates in Information Theory -> surprise in bits:

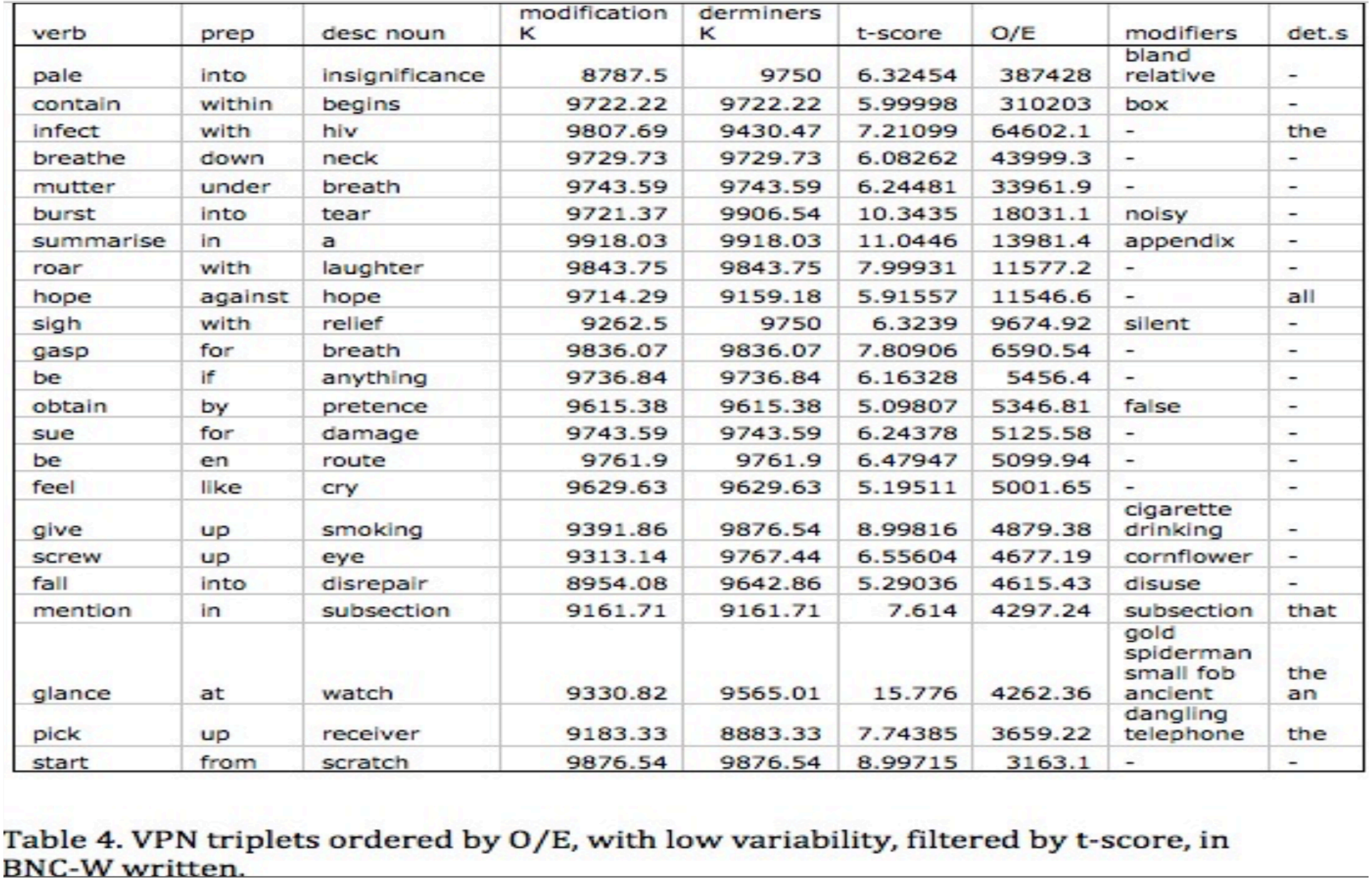
$$MI(x;y) = \log_2 \frac{P(x,y)}{P(x) * P(y)}$$

- O/E simply divides Observation by Expectation:

$$O/E = \frac{P(x,y)}{P(x) * P(y)} = \frac{f(x,y) * N * N}{N * f(x) * f(y)} = \frac{f(x,y) * N}{f(x) * f(y)}$$

2.1 O/E (Observed divided by Expected, O over E)

- Applied to verb-PP constructions in the BNC (Lehmann & Schneider 2011)



verb	prep	desc noun	modification K	derminers K	t-score	O/E	modifiers	det.s
pale	into	insignificance	8787.5	9750	6.32454	387428	bland relative	-
contain	within	begins	9722.22	9722.22	5.99998	310203	box	-
infect	with	hiv	9807.69	9430.47	7.21099	64602.1	-	the
breathe	down	neck	9729.73	9729.73	6.08262	43999.3	-	-
mutter	under	breath	9743.59	9743.59	6.24481	33961.9	-	-
burst	into	tear	9721.37	9906.54	10.3435	18031.1	noisy	-
summarise	in	a	9918.03	9918.03	11.0446	13981.4	appendix	-
roar	with	laughter	9843.75	9843.75	7.99931	11577.2	-	-
hope	against	hope	9714.29	9159.18	5.91557	11546.6	-	all
sigh	with	relief	9262.5	9750	6.3239	9674.92	silent	-
gasp	for	breath	9836.07	9836.07	7.80906	6590.54	-	-
be	if	anything	9736.84	9736.84	6.16328	5456.4	-	-
obtain	by	pretence	9615.38	9615.38	5.09807	5346.81	false	-
sue	for	damage	9743.59	9743.59	6.24378	5125.58	-	-
be	en	route	9761.9	9761.9	6.47947	5099.94	-	-
feel	like	cry	9629.63	9629.63	5.19511	5001.65	-	-
give	up	smoking	9391.86	9876.54	8.99816	4879.38	cigarette drinking	-
screw	up	eye	9313.14	9767.44	6.55604	4677.19	cornflower	-
fall	into	disrepair	8954.08	9642.86	5.29036	4615.43	disuse	-
mention	in	subsection	9161.71	9161.71	7.614	4297.24	subsection	that
glance	at	watch	9330.82	9565.01	15.776	4262.36	gold spiderman small fob ancient	the an
pick	up	receiver	9183.33	8883.33	7.74385	3659.22	dangling telephone	the
start	from	scratch	9876.54	9876.54	8.99715	3163.1	-	-

Table 4. VPN triplets ordered by O/E, with low variability, filtered by t-score, in BNC-W written.

2.2 Collocation Ratio

For detecting L2 errors and innovations we want to detect verb-PP combinations which

1) are frequent enough to reach collocation status
2) are collocations in L2
3) but not, or much less so, in L1

If we apply traditional collocation measures we fail to see 3)

A successful measure for 3) is the collocation ratio (Schneider and Zipp 2013): if $c_{\mathrm{L}1}(a,b)$ is a collocation measure c for L1 of words w_1 and w_2, then:

$$\text{Collocation ratio} = c_{\mathrm{L}2}(w_1,w_2) \,/\, c_{\mathrm{L}1}(w_1,w_2)$$

It is a measure of overuse, of „overcollocability“, a meta O/E measure

2.2 Collocation Ratio with O/E (=Observed / Expected)

We consider verb-PP combinations:

w_1=verb or adjective, w_2=preposition or verbal particle

As L1 corpus we use the BNC, as L2 ICLE

When using the **collocation measure O/E** the ratio is

$$O/E\ ratio = \frac{O/E(ICLE)}{O/E(BNC)} = \frac{\frac{O(ICLE)}{E(ICLE)}}{\frac{O(BNC)}{E(BNC)}} = \frac{\frac{O_{ICLE}(R,w_1,w_2)\cdot N_{ICLE}}{O_{ICLE}(R,w_1)\cdot O_{ICLE}(R,w_2)}}{\frac{O_{BNC}(R,w_1,w_2)\cdot N_{BNC}}{O_{BNC}(R,w_1)\cdot O_{BNC}(R,w_2)}}$$

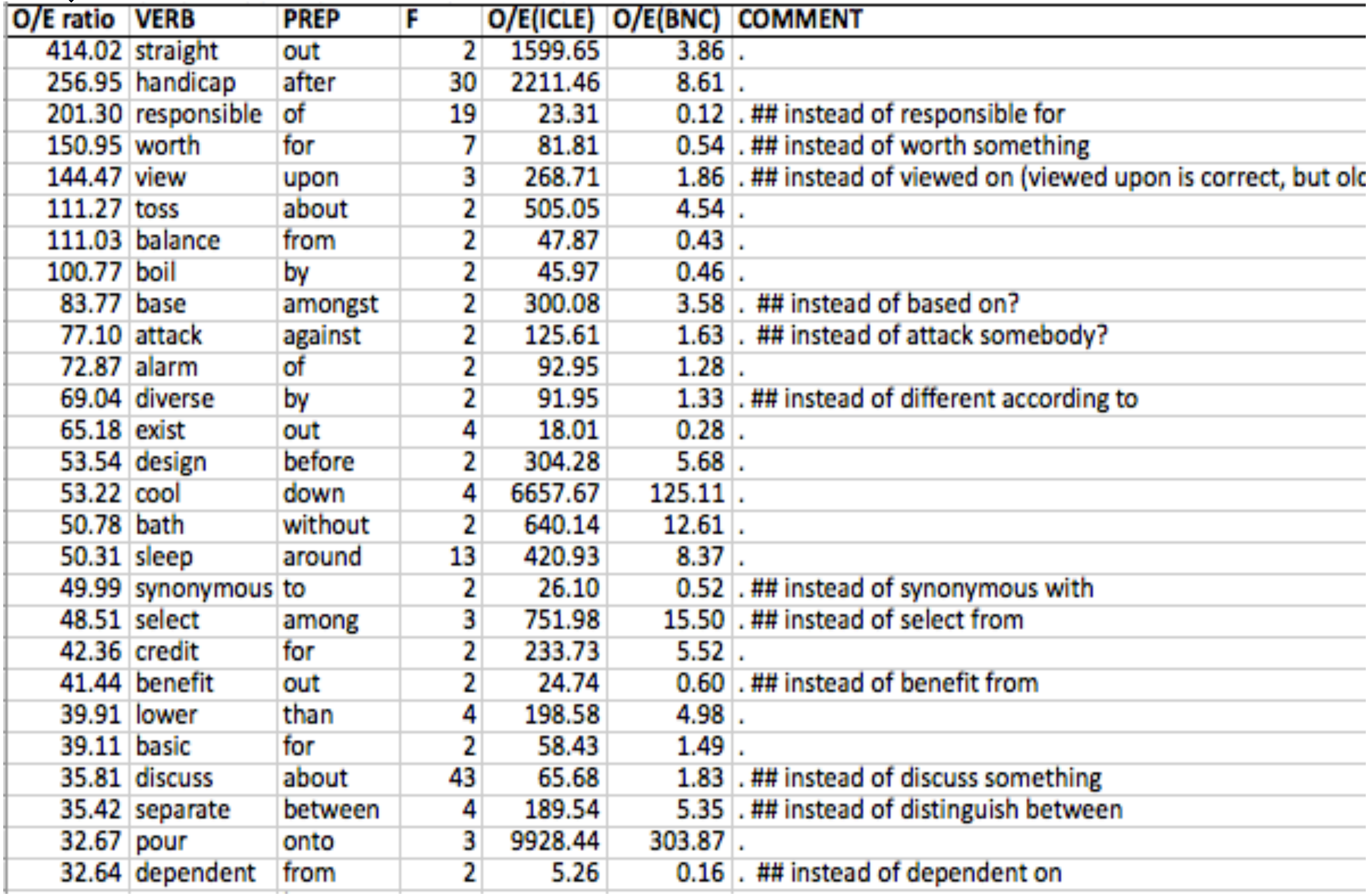
This is itself an O/E measure: O = O/E(ICLE); E = O/E(BNC)

For the T-Score collo. a formulation in terms of O and E (Evert 2009) is:

$$T = \frac{O-E}{\sqrt{(O)}} \to T\ ratio = \frac{T(ICLE)}{T(BNC)} = \frac{\frac{O(ICLE)-E(ICLE)}{\sqrt{O(ICLE)}}}{\frac{O(BNC)-E(BNC)}{\sqrt{O(BNC)}}}$$

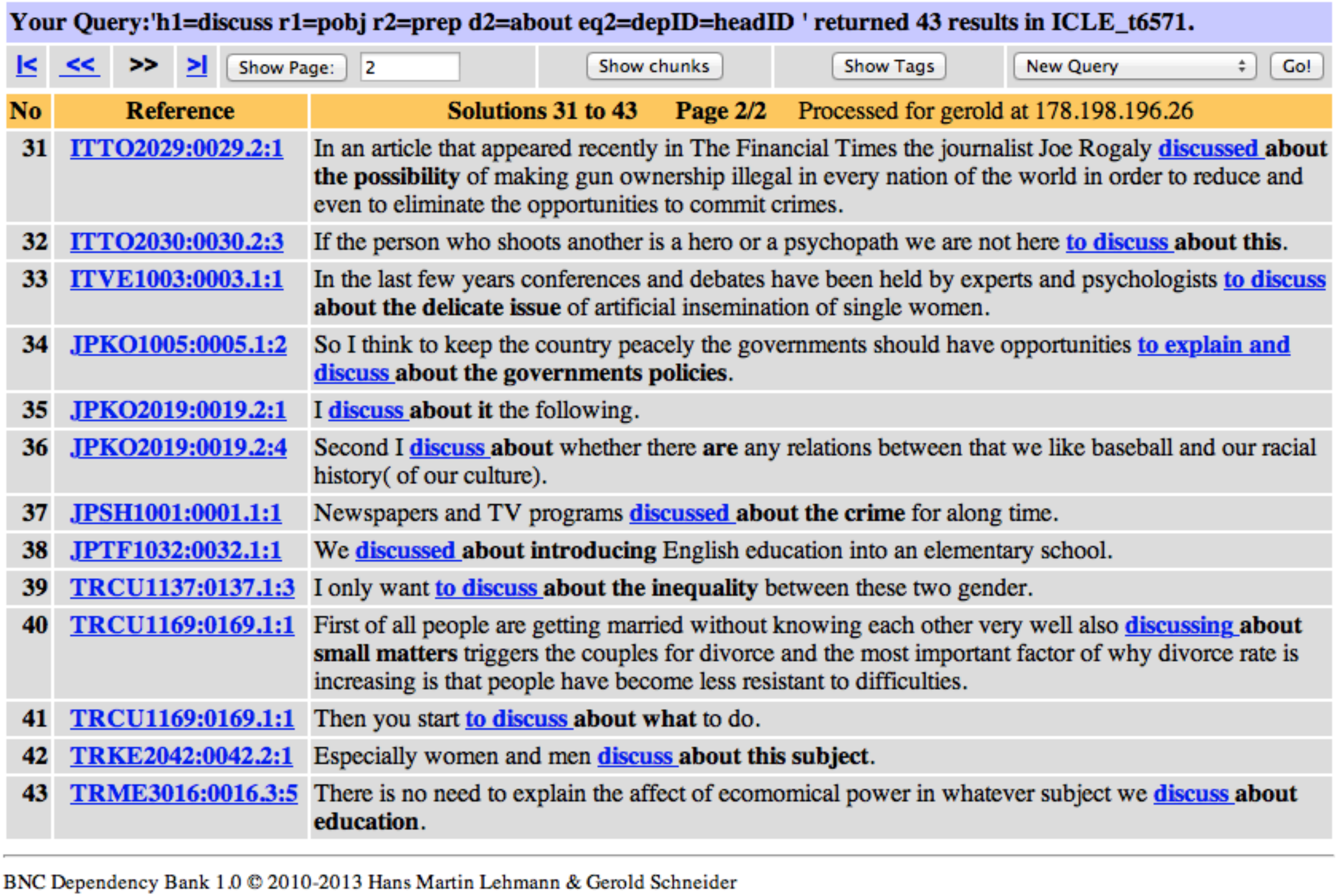
-sort

2.3 Data-driven verb-PP: O/E results



O/E ratio	VERB	PREP	F	O/E(ICLE)	O/E(BNC)	COMMENT
414.02	straight	out	2	1599.65	3.86	.
256.95	handicap	after	30	2211.46	8.61	.
201.30	responsible	of	19	23.31	0.12	. ## instead of responsible for
150.95	worth	for	7	81.81	0.54	. ## instead of worth something
144.47	view	upon	3	268.71	1.86	. ## instead of viewed on (viewed upon is correct, but old
111.27	toss	about	2	505.05	4.54	.
111.03	balance	from	2	47.87	0.43	.
100.77	boil	by	2	45.97	0.46	.
83.77	base	amongst	2	300.08	3.58	. ## instead of based on?
77.10	attack	against	2	125.61	1.63	. ## instead of attack somebody?
72.87	alarm	of	2	92.95	1.28	.
69.04	diverse	by	2	91.95	1.33	. ## instead of different according to
65.18	exist	out	4	18.01	0.28	.
53.54	design	before	2	304.28	5.68	.
53.22	cool	down	4	6657.67	125.11	.
50.78	bath	without	2	640.14	12.61	.
50.31	sleep	around	13	420.93	8.37	.
49.99	synonymous	to	2	26.10	0.52	. ## instead of synonymous with
48.51	select	among	3	751.98	15.50	. ## instead of select from
42.36	credit	for	2	233.73	5.52	.
41.44	benefit	out	2	24.74	0.60	. ## instead of benefit from
39.91	lower	than	4	198.58	4.98	.
39.11	basic	for	2	58.43	1.49	.
35.81	discuss	about	43	65.68	1.83	. ## instead of discuss something
35.42	separate	between	4	189.54	5.35	. ## instead of distinguish between
32.67	pour	onto	3	9928.44	303.87	.
32.64	dependent	from	2	5.26	0.16	. ## instead of dependent on

2.3. Data-driven verb-PP: O/E examples



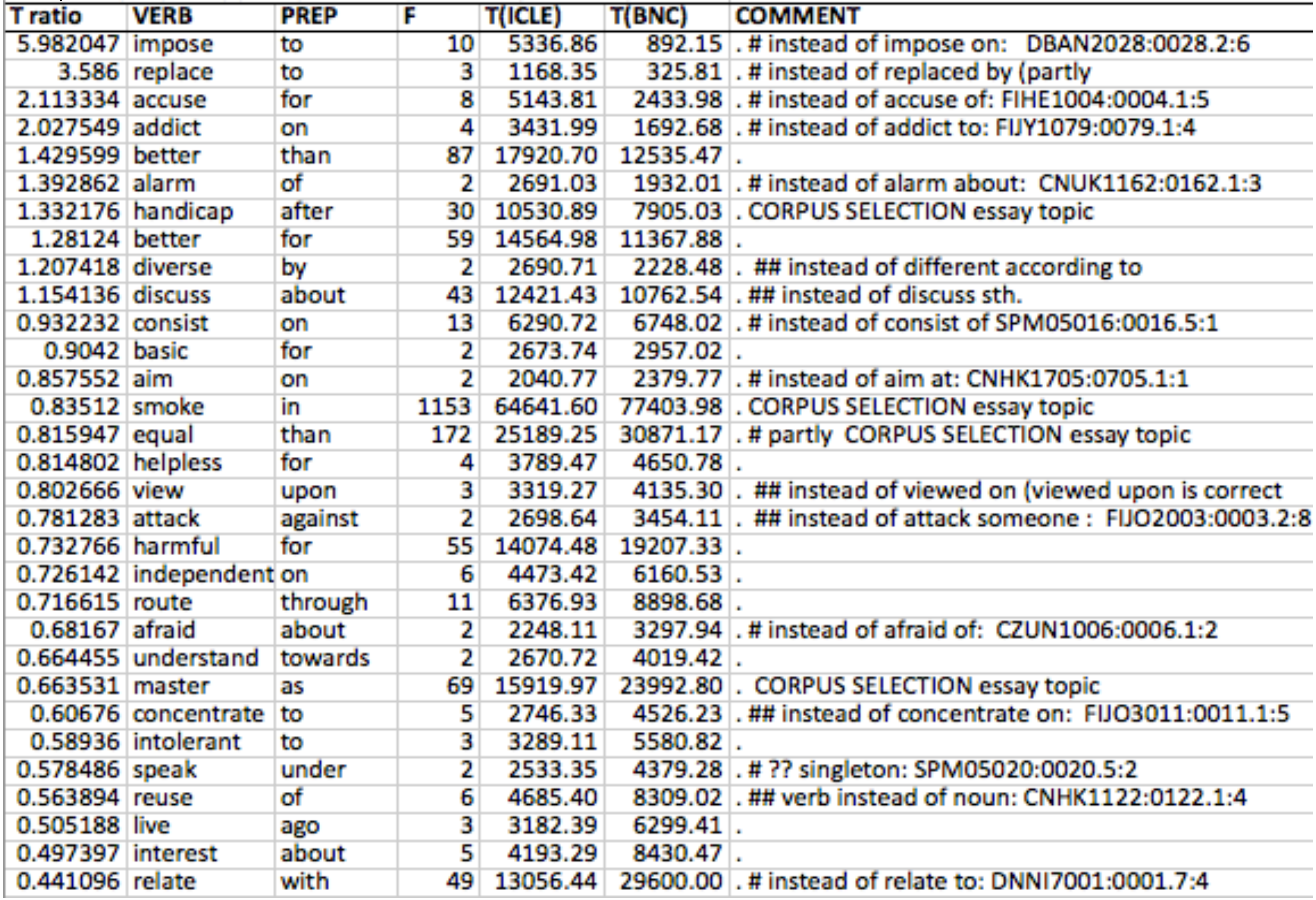
Your Query:'h1=discuss r1=pobj r2=prep d2=about eq2=depID=headID ' returned 43 results in ICLE_t6571.

No	Reference	Solutions 31 to 43 Page 2/2 Processed for gerold at 178.198.196.26
31	ITTO2029:0029.2:1	In an article that appeared recently in The Financial Times the journalist Joe Rogaly discussed **about the possibility** of making gun ownership illegal in every nation of the world in order to reduce and even to eliminate the opportunities to commit crimes.
32	ITTO2030:0030.2:3	If the person who shoots another is a hero or a psychopath we are not here to discuss **about this**.
33	ITVE1003:0003.1:1	In the last few years conferences and debates have been held by experts and psychologists to discuss **about the delicate issue** of artificial insemination of single women.
34	JPKO1005:0005.1:2	So I think to keep the country peacely the governments should have opportunities to explain and discuss **about the governments policies**.
35	JPKO2019:0019.2:1	I discuss **about it** the following.
36	JPKO2019:0019.2:4	Second I discuss **about** whether there **are** any relations between that we like baseball and our racial history(of our culture).
37	JPSH1001:0001.1:1	Newspapers and TV programs discussed **about the crime** for along time.
38	JPTF1032:0032.1:1	We discussed **about introducing** English education into an elementary school.
39	TRCU1137:0137.1:3	I only want to discuss **about the inequality** between these two gender.
40	TRCU1169:0169.1:1	First of all people are getting married without knowing each other very well also discussing **about small matters** triggers the couples for divorce and the most important factor of why divorce rate is increasing is that people have become less resistant to difficulties.
41	TRCU1169:0169.1:1	Then you start to discuss **about what** to do.
42	TRKE2042:0042.2:1	Especially women and men discuss **about this subject**.
43	TRME3016:0016.3:5	There is no need to explain the affect of ecomomical power in whatever subject we discuss **about education**.

BNC Dependency Bank 1.0 © 2010-2013 Hans Martin Lehmann & Gerold Schneider

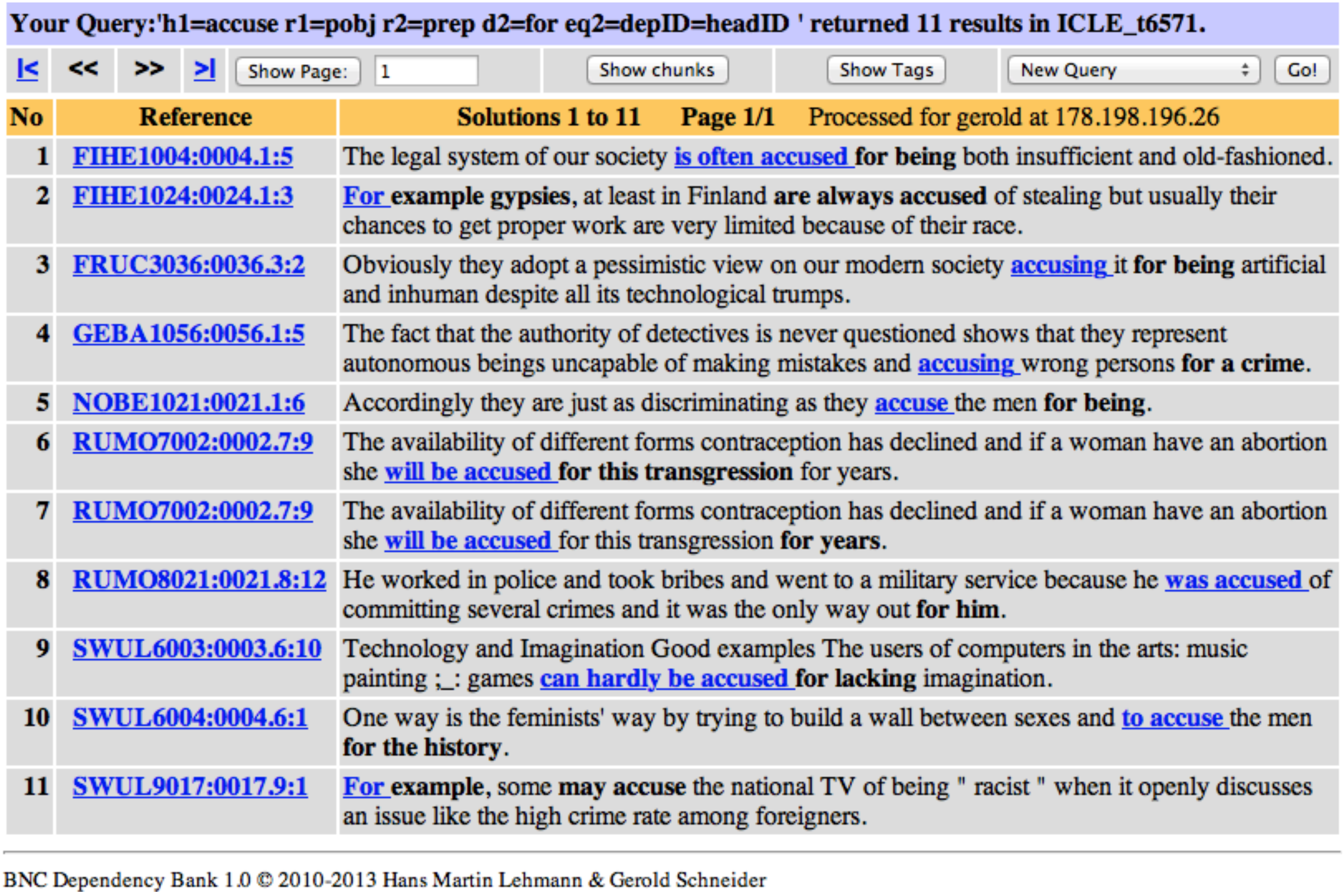
-sort

2.3. Data-driven verb-PP: T-score results & example



T ratio	VERB	PREP	F	T(ICLE)	T(BNC)	COMMENT
5.982047	impose	to	10	5336.86	892.15	. # instead of impose on: DBAN2028:0028.2:6
3.586	replace	to	3	1168.35	325.81	. # instead of replaced by (partly
2.113334	accuse	for	8	5143.81	2433.98	. # instead of accuse of: FIHE1004:0004.1:5
2.027549	addict	on	4	3431.99	1692.68	. # instead of addict to: FIJY1079:0079.1:4
1.429599	better	than	87	17920.70	12535.47	.
1.392862	alarm	of	2	2691.03	1932.01	. # instead of alarm about: CNUK1162:0162.1:3
1.332176	handicap	after	30	10530.89	7905.03	. CORPUS SELECTION essay topic
1.28124	better	for	59	14564.98	11367.88	.
1.207418	diverse	by	2	2690.71	2228.48	. ## instead of different according to
1.154136	discuss	about	43	12421.43	10762.54	. ## instead of discuss sth.
0.932232	consist	on	13	6290.72	6748.02	. # instead of consist of SPM05016:0016.5:1
0.9042	basic	for	2	2673.74	2957.02	.
0.857552	aim	on	2	2040.77	2379.77	. # instead of aim at: CNHK1705:0705.1:1
0.83512	smoke	in	1153	64641.60	77403.98	. CORPUS SELECTION essay topic
0.815947	equal	than	172	25189.25	30871.17	. # partly CORPUS SELECTION essay topic
0.814802	helpless	for	4	3789.47	4650.78	.
0.802666	view	upon	3	3319.27	4135.30	. ## instead of viewed on (viewed upon is correct
0.781283	attack	against	2	2698.64	3454.11	. ## instead of attack someone : FIJO2003:0003.2:8
0.732766	harmful	for	55	14074.48	19207.33	.
0.726142	independent	on	6	4473.42	6160.53	.
0.716615	route	through	11	6376.93	8898.68	.
0.68167	afraid	about	2	2248.11	3297.94	. # instead of afraid of: CZUN1006:0006.1:2
0.664455	understand	towards	2	2670.72	4019.42	.
0.663531	master	as	69	15919.97	23992.80	. CORPUS SELECTION essay topic
0.60676	concentrate	to	5	2746.33	4526.23	. ## instead of concentrate on: FIJO3011:0011.1:5
0.58936	intolerant	to	3	3289.11	5580.82	.
0.578486	speak	under	2	2533.35	4379.28	. # ?? singleton: SPM05020:0020.5:2
0.563894	reuse	of	6	4685.40	8309.02	. ## verb instead of noun: CNHK1122:0122.1:4
0.505188	live	ago	3	3182.39	6299.41	.
0.497397	interest	about	5	4193.29	8430.47	.
0.441096	relate	with	49	13056.44	29600.00	. # instead of relate to: DNNI7001:0001.7:4

2.3. Data-driven verb-PP: T-score results & example



Your Query:'h1=accuse r1=pobj r2=prep d2=for eq2=depID=headID ' returned 11 results in ICLE_t6571.

No	Reference	Solutions 1 to 11 Page 1/1 Processed for gerold at 178.198.196.26
1	FIHE1004:0004.1:5	The legal system of our society is often accused **for being** both insufficient and old-fashioned.
2	FIHE1024:0024.1:3	For **example gypsies**, at least in Finland **are always accused** of stealing but usually their chances to get proper work are very limited because of their race.
3	FRUC3036:0036.3:2	Obviously they adopt a pessimistic view on our modern society accusing it **for being** artificial and inhuman despite all its technological trumps.
4	GEBA1056:0056.1:5	The fact that the authority of detectives is never questioned shows that they represent autonomous beings uncapable of making mistakes and accusing wrong persons **for a crime**.
5	NOBE1021:0021.1:6	Accordingly they are just as discriminating as they accuse the men **for being**.
6	RUMO7002:0002.7:9	The availability of different forms contraception has declined and if a woman have an abortion she will be accused **for this transgression** for years.
7	RUMO7002:0002.7:9	The availability of different forms contraception has declined and if a woman have an abortion she will be accused for this transgression **for years**.
8	RUMO8021:0021.8:12	He worked in police and took bribes and went to a military service because he was accused of committing several crimes and it was the only way out **for him**.
9	SWUL6003:0003.6:10	Technology and Imagination Good examples The users of computers in the arts: music painting ;_: games can hardly be accused **for lacking** imagination.
10	SWUL6004:0004.6:1	One way is the feminists' way by trying to build a wall between sexes and to accuse the men **for the history**.
11	SWUL9017:0017.9:1	For **example**, some **may accuse** the national TV of being " racist " when it openly discusses an issue like the high crime rate among foreigners.

BNC Dependency Bank 1.0 © 2010-2013 Hans Martin Lehmann & Gerold Schneider

2.3. Data-driven verb-PP: Evaluation, Precision

Evaluation:
P=12/30 = 40%
P=20/60 = 33%

For Text Mining experts, this seems modest.

But manual filtering based on inspecting the hits is quite simple.

We could also increase precision by setting a filter on O/E(BNC) corresponding to the criterion that innovations/errors should not have high collocational status in the native variant.

If we set a filter of O/E(BNC)<5, precision rises to above 50%, but at the trade-off of lower recall: e.g. *select among* and *separate between* would not be returned

2.4. Data-driven verb-PP: negative collo or unseen in BNC

The combinations which have negative collocation in BNC are boundless.

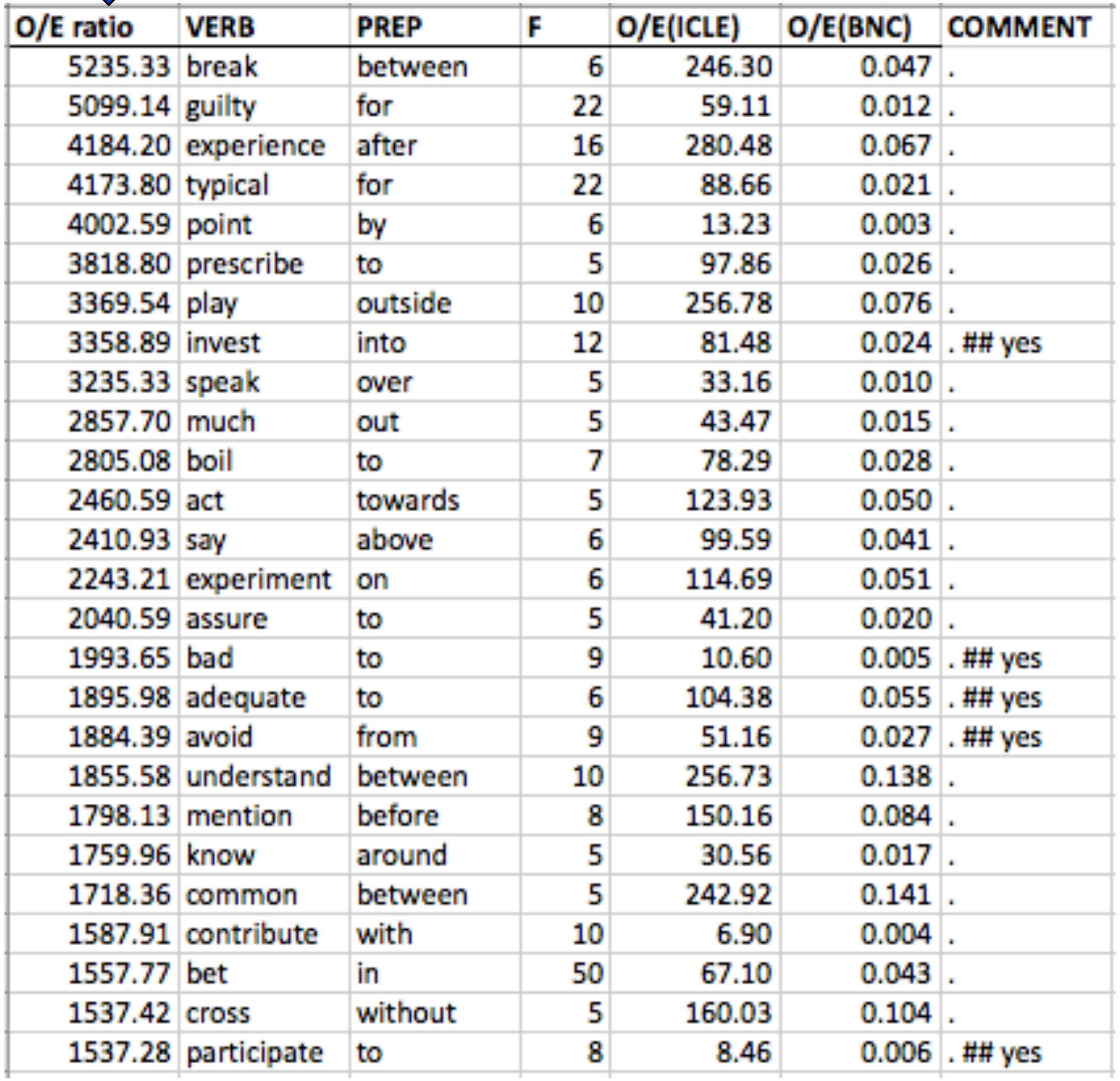
Here: f > 4, negative collocation ==

Most candidates which are not present (unseen) in the BNC

- *could* also appear there: sparse data
- or are parsing errors

Some frequent ones, however, are innovations.
This is an abundant resource with hundreds of candidates, but quite low precision.
(next slide)

-sort



O/E ratio	VERB	PREP	F	O/E(ICLE)	O/E(BNC)	COMMENT
5235.33	break	between	6	246.30	0.047	.
5099.14	guilty	for	22	59.11	0.012	.
4184.20	experience	after	16	280.48	0.067	.
4173.80	typical	for	22	88.66	0.021	.
4002.59	point	by	6	13.23	0.003	.
3818.80	prescribe	to	5	97.86	0.026	.
3369.54	play	outside	10	256.78	0.076	.
3358.89	invest	into	12	81.48	0.024	. ## yes
3235.33	speak	over	5	33.16	0.010	.
2857.70	much	out	5	43.47	0.015	.
2805.08	boil	to	7	78.29	0.028	.
2460.59	act	towards	5	123.93	0.050	.
2410.93	say	above	6	99.59	0.041	.
2243.21	experiment	on	6	114.69	0.051	.
2040.59	assure	to	5	41.20	0.020	.
1993.65	bad	to	9	10.60	0.005	. ## yes
1895.98	adequate	to	6	104.38	0.055	. ## yes
1884.39	avoid	from	9	51.16	0.027	. ## yes
1855.58	understand	between	10	256.73	0.138	.
1798.13	mention	before	8	150.16	0.084	.
1759.96	know	around	5	30.56	0.017	.
1718.36	common	between	5	242.92	0.141	.
1587.91	contribute	with	10	6.90	0.004	.
1557.77	bet	in	50	67.10	0.043	.
1537.42	cross	without	5	160.03	0.104	.
1537.28	participate	to	8	8.46	0.006	. ## yes

2.4

* filter of O/E(BNC) <5,

* added a smoothing count of 0.5 (new fifth column) to types unseen in BNC.

Note: many semantic preps instead of functional preps.

O/E ratio	VERB/ADJ.	PREP	F(ICLE)	F(BNC)	O/E(ICLE)	O/E(BNC)	COMMENT
488.81	critical	towards	7	0.5	1511.26	3.09	instead of *critical to*
201.30	responsible	of	19	2	23.31	0.12	instead of *responsible for*
189.01	critical	against	4	0.5	370.22	1.96	instead of *critical to*
150.95	worth	for	7	1	81.81	0.54	instead of *worth something*
145.67	superior	than	22	0.5	434.65	2.98	instead of *superior to*
138.75	indulge	into	6	0.5	61.11	0.44	instead of *indulge in*
110.11	overcrowd	at	32	0.5	485.00	4.40	CORPUS essay topic
69.11	destructive	for	5	1	166.95	2.42	instead of *destructive to*
65.18	exist	out	4	2	18.01	0.28	
39.91	lower	than	4	2	198.58	4.98	
35.81	discuss	about	43	7	65.68	1.83	instead of *discuss something*
34.27	conscious	about	10	2	124.19	3.62	instead of *conscious of*
32.06	helpless	for	4	1	66.78	2.08	
31.55	possible	out	4	5	30.37	0.96	
30.60	recur	to	4	7	125.26	4.09	
29.94	dependent	of	8	4	19.34	0.65	instead of *dependent on*
24.63	belong	into	4	2	6.63	0.27	instead of *belong to*
23.59	renounce	to	9	3	108.40	4.60	
23.07	decide	over	7	13	102.14	4.43	CORPUS essay topic
21.96	inherent	to	9	13	78.29	3.56	
20.46	relate	with	49	76	32.98	1.61	instead of *relate to*
19.80	aware	about	4	1	5.94	0.30	instead of *aware of*
19.67	aspire	for	4	3	51.94	2.64	instead of *aspire to*
18.21	guilty	for	22	28	59.11	3.25	instead of *guilty of*
17.72	little	by	11	36	70.80	4.00	
17.67	produce	out	4	30	44.85	2.54	
17.19	accuse	for	8	19	18.33	1.07	instead of *accuse of*
15.39	interest	to	7	0.5	11.54	0.75	
15.01	specialize	on	4	4	40.24	2.68	
15.01	deal	about	4	2	3.91	0.26	instead of *deal with*

2.5. Which metric? O/E vs t-score

Some combinations are detected by both O/E and t-score

e.g. *basic for, discuss about, helpless for, relate with*

But each measure brings up its own (relevant) combinations, including different prepositions with identical verbs/adjectives

cf. *independent from* (O/E) – *independent on* (T)

It is therefore useful to combine the two measures. In our collocation ratio measure, the different characteristics of the metrics are less clearly apparent.

2.6 Cognitive origin of novel combinations

Standard combination	Novel combination	Possible origin: L1 transfer / analogy
To discuss sth To attack sb To be credited with To relate to	To discuss about sth To attack against sb To be credited for To relate with	Discussion about Attack_NN against Credit_NN for Relations with
Independent of	Independent on	Dependent on
To separate sth from sth To be viewed as To arrive at Content with Afraid of	To separate between To be viewed upon as To arrive to Content about Afraid about	To distinguish between To be looked upon as To get to Happy with/about Scared about
Inherent in Select from	Inherent to Select among	FR. inhérent à DE. Auswählen zwischen

2.7. Differences between EFL and ESL

The relationship between ESL (second language) and EFL (foreign language) has moved into research focus (e.g. Nesselhauf 2009). It is hard to claim that similar phenomena are innovations in ESL but errors in EFL.

We have so far compared to BNC=L1 as reference corpus. We can apply the same approach to find differences between EFL and ESL: EFL as application corpus, compared to ESL as reference corpus.

- We ran a version with particularly strict O/E(ICE 5 ESL)<2, counting unseen instances as 0.5, aiming at a core set of typical verb/adjective + preposition innovations which only EFL speakers but not ESL speakers use (next slide)
- Noun-analogies (noun complementation patterns taken over to the verb) are very rare (only one, *assist to*) compared to ESL
- preposition *to* seems to be used too generically: 7 out of the 13 true positives involve *to*. There might be a trend to use *to* as a generic marker for indirect objects, particularly in Romance langs

2.7. Differences between EFL and ESL: EFL, not ESL

O/E ratio	VERB/ADJ	PREP	F(ICLE)	F(ICE-5 ESL)	O/E(ICLE)	O/E(ICE-5 ESL)	COMMENT
35.97	equivalent	in	5	0.5	35.34	0.98	
34.19	assist	to	6	1	27.63	0.81	instead of *assist sth.*
25.68	accuse	for	8	0.5	18.33	0.71	instead of *accuse of*
22.29	wrong	at	6	0.5	24.38	1.09	
21.61	explain	from	8	0.5	16.03	0.74	
21.28	stay	like	5	0.5	13.53	0.64	
15.45	participate	to	8	1	8.46	0.55	instead of *participate in*
14.10	arise	by	6	0.5	12.14	0.86	instead of *due to/from*
12.60	employ	of	5	0.5	18.19	1.44	parsing error
11.35	benefit	to	13	1	10.49	0.92	instead of *be of benefit to*
9.10	impose	to	10	1	8.15	0.90	instead of *impose on*
8.06	oppose	in	6	0.5	5.05	0.63	
5.63	equal	for	9	0.5	4.22	0.75	instead of *equal to*
5.51	discuss	of	5	0.5	4.22	0.77	
5.40	remain	to	5	2	4.33	0.80	
5.34	necessary	with	6	0.5	6.70	1.25	instead of *necessary for*
5.08	keep	into	5	1	4.22	0.83	instead of *keep at*
5.05	reflect	to	5	1	5.12	1.01	instead of *reflect sth.*
4.95	confront	to	6	0.5	7.17	1.45	instead of *confront with*
4.93	discuss	for	13	2	6.13	1.24	
4.72	popular	to	6	0.5	4.84	1.03	instead of *popular for*

2.7. Innovation vs. Error

- Deletion of Hapax Legomena cuts out some obvious errors (misproductions)
- Recurrence=Systematicity can be covered quite well by using collocation measures
- ESL / ENL (innovation): analogy to the complementation patterns of nouns seems particularly frequent among ESL speakers
- EFL / ESL (error): preposition *to* is used too generically
- EFL / ESL tells us which new patterns are particularly **different**
- We can also use a method telling us which are particularly **similar**: Detect EFL / ENL , but only report those which have *similar* O/E ratio: As a threshold we set that O/E(ICLE) is maximally 3 times larger than O/E(ICE) or vice versa:

2.7. Innovation or Error: ESL & EFL are similar

O/E ratio	VERB/ADJ.	PREP	F(ICLE)	O/E (ICLE)	O/E (ICE-5 ESL)	O/E (BNC)	COMMENT
145.67	superior	than	22	434.6	565.61	2.98	instead of *superior to*
138.75	indulge	into	6	61.11	28.10	0.44	instead of *indulge in*
35.81	discuss	about	43	65.68	83.59	1.83	instead of *discuss sth.*
34.27	conscious	about	10	124.1	78.30	3.62	instead of *conscious of*
19.67	aspire	for	4	51.94	31.93	2.64	instead of *aspire to*
17.72	little	by	11	70.80	38.50	4.00	
15.39	interest	to	7	11.54	6.08	0.75	
14.29	point	by	6	13.23	5.57	0.93	
13.49	commensurate	to	4	22.37	49.29	1.66	
13.24	interest	for	26	63.97	41.70	4.83	
12.94	speak	over	5	33.16	13.06	2.56	
10.65	own	to	8	23.20	8.80	2.18	instead of *owing to* (partly)
10.28	watch	than	4	17.52	18.76	1.70	
9.75	capable	in	5	2.83	2.97	0.29	instead of *capable of/to*
9.10	deprive	from	10	18.64	12.64	2.05	
8.84	study	about	8	11.66	26.05	1.32	instead of *study sth.*
8.62	charge	of	4	30.98	11.88	3.59	instead of *change sth*/noun
7.86	shut	to	7	36.53	27.73	4.65	
7.28	face	to	35	19.64	7.86	2.70	instead of *face sth.*
7.24	state	about	4	25.04	11.77	3.46	
6.81	invest	to	5	5.44	2.93	0.80	instead of *invest in*
6.66	speed	in	5	33.13	27.33	4.98	
6.65	waste	for	8	24.28	18.73	3.65	
6.52	reward	to	6	18.07	24.65	2.77	
6.37	associate	to	4	3.89	3.29	0.61	instead of *associate with*
6.36	strike	to	6	16.48	6.16	2.59	
6.02	know	over	4	16.60	9.30	2.76	
5.95	afford	with	4	18.63	33.91	3.13	
5.89	steal	to	6	9.39	3.21	1.59	instead of *steal from*
5.88	sum	in	4	22.32	30.50	3.80	
5.51	influence	on	15	15.21	6.40	2.76	instead of noun(partly)
5.30	depend	from	9	4.84	1.76	0.91	instead of *depend on*
5.19	search	from	5	15.06	7.52	2.90	instead of *search on*

3. Helping Learners with non-compositional items

- Verb-PP structures are one area of non-compositionality
- Non-compositionality is generally hard to learn, except in closely related languages, where some idioms, collocations, lexical preferences etc. are similar
- We use parallel corpora, detecting translation that are hard:

  Collocations are non-compositional and different form the speaker’s native language

  - Adjective-noun combinations
  - Verb-object combinations (light verbs)
  - Verb-PP constructions

3. Helping Learners with non-compositional items

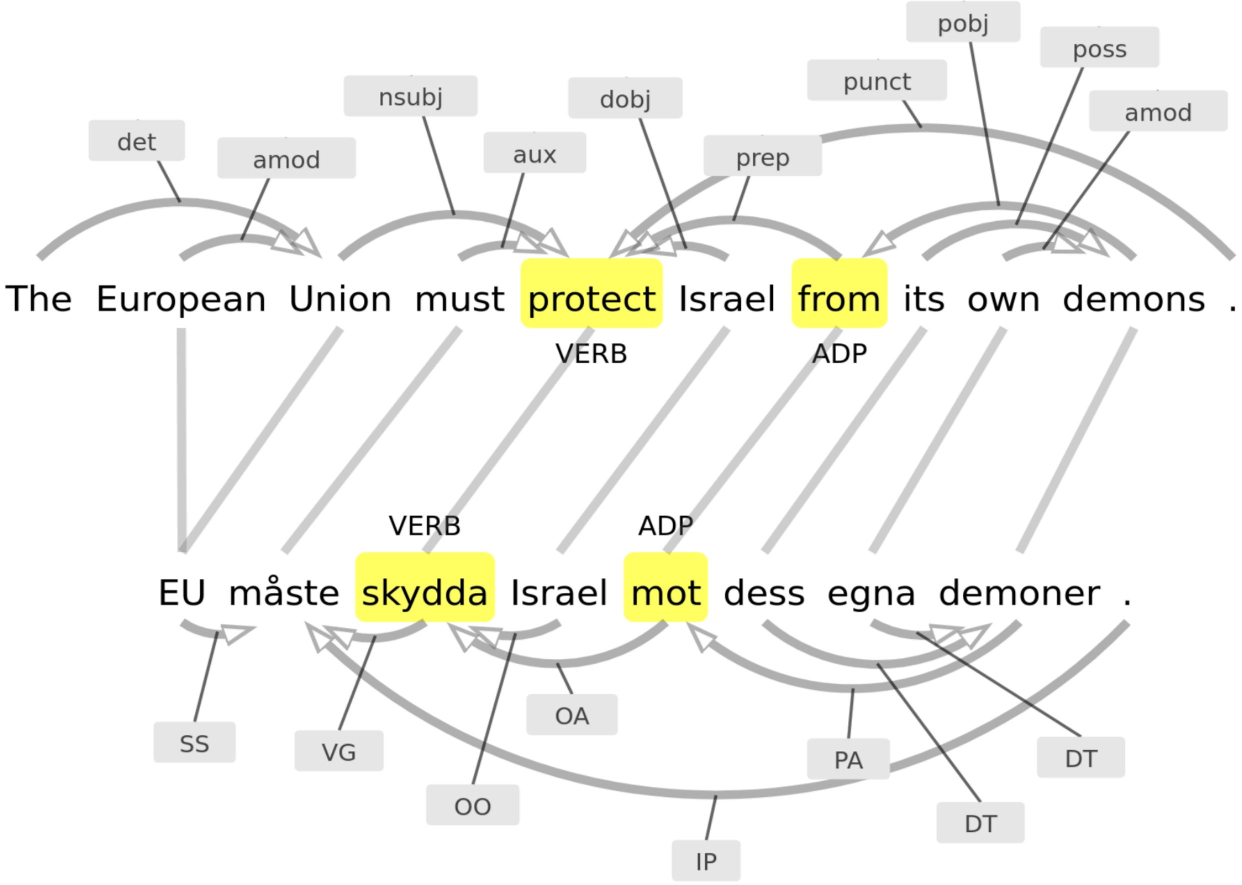
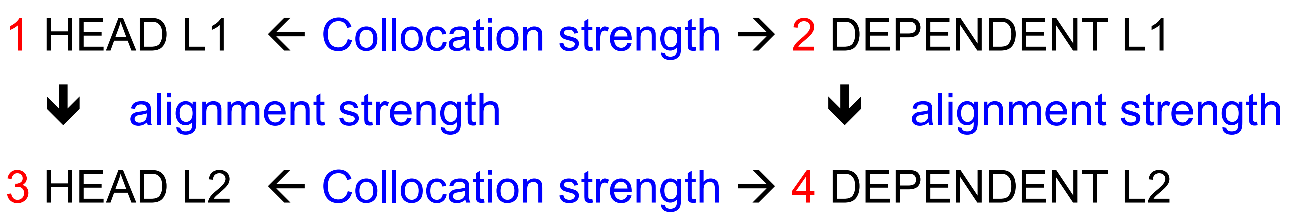

pobj
poss
punct
amod
nsubj
dobj
det
amod
aux
prep
The European Union must protect Israel from its own demons .
VERB
ADP
VERB
ADP
EU måste skydda Israel mot dess egna demoner .
SS
VG
OO
OA
PA
DT
DT
IP

Figure 1: A constellation consisting of two aligned verbs with corresponding aligned prepositions.

3. Helping Learners with non-compositional items



1 HEAD L1 ← Collocation strength → 2 DEPENDENT L1

↓ alignment strength ↓ alignment strength

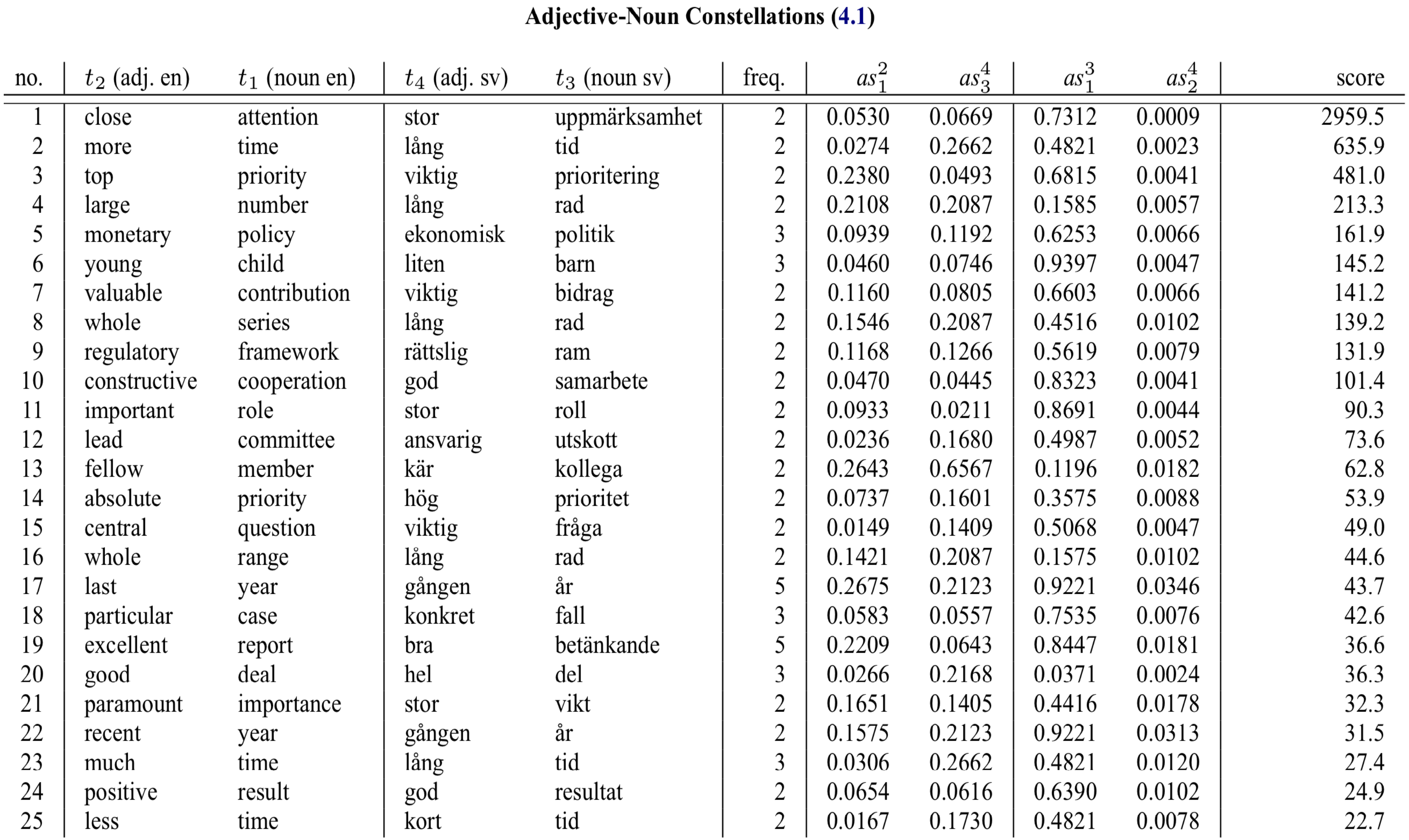
3 HEAD L2 ← Collocation strength → 4 DEPENDENT L2

- Direct & frequent translations have high alignment strength: as13 and as24
- We can use collocation measures for the alignment strength (t-score, z-score, O/E, MI, etc.)
- Non-compositional idioms have
  - high collocation strength in both languages: as12 and as34
  - high alignment strength on the head: as13
  - **low** alignment strength on the dependent: as24
    → unusual, we are looking for non-direct translations

→ score=as12*as34*as13/as24

→ score=as12*as34*as13/as24*log(c) ## frequency-weighted version

3. Helping Learners with non-compositional items

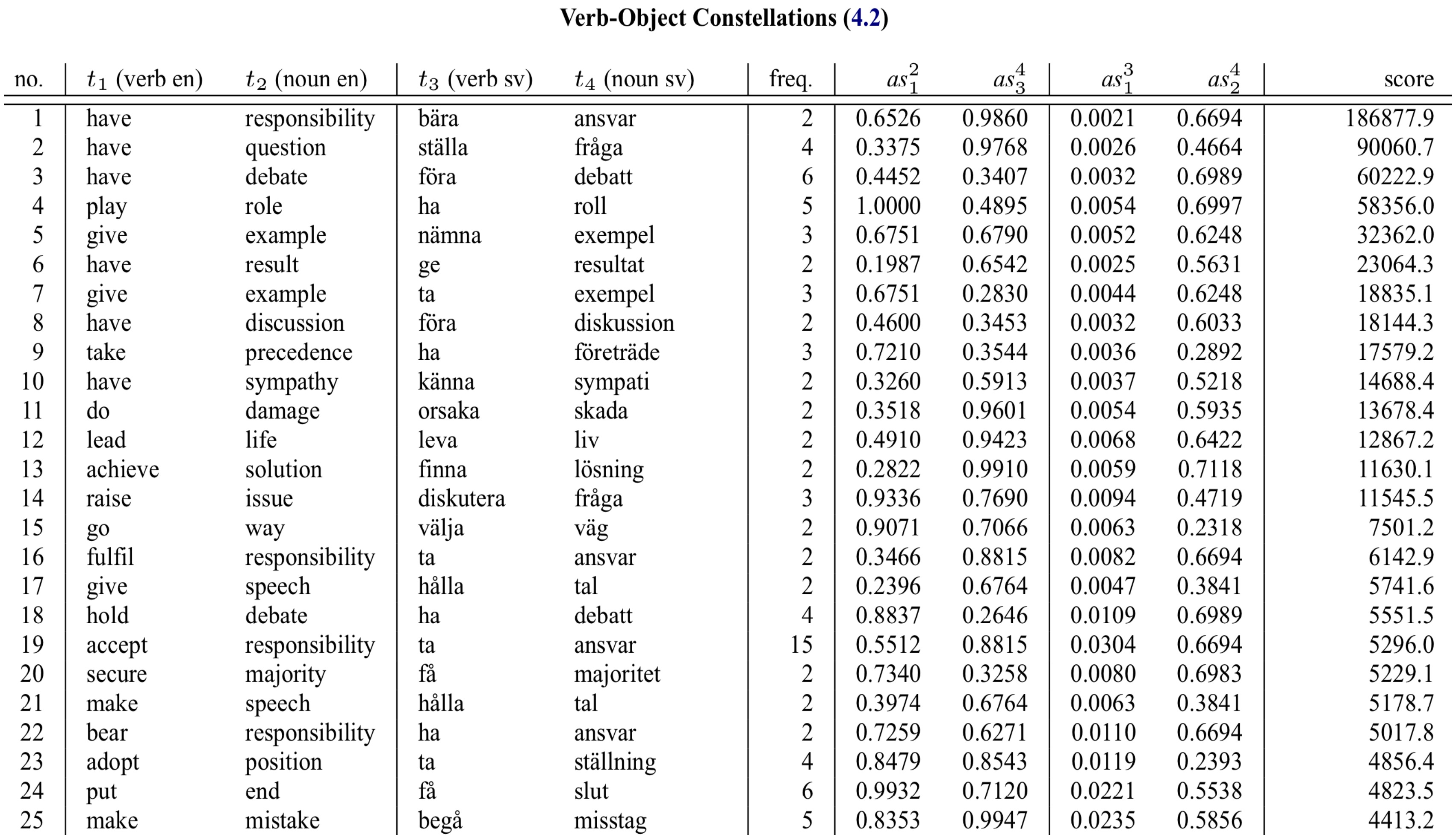


Adjective-Noun Constellations (4.1)

no.	t_2 (adj. en)	t_1 (noun en)	t_4 (adj. sv)	t_3 (noun sv)	freq.	as_1^2	as_3^4	as_1^3	as_2^4	score
1	close	attention	stor	uppmärksamhet	2	0.0530	0.0669	0.7312	0.0009	2959.5
2	more	time	lång	tid	2	0.0274	0.2662	0.4821	0.0023	635.9
3	top	priority	viktig	prioritering	2	0.2380	0.0493	0.6815	0.0041	481.0
4	large	number	lång	rad	2	0.2108	0.2087	0.1585	0.0057	213.3
5	monetary	policy	ekonomisk	politik	3	0.0939	0.1192	0.6253	0.0066	161.9
6	young	child	liten	barn	3	0.0460	0.0746	0.9397	0.0047	145.2
7	valuable	contribution	viktig	bidrag	2	0.1160	0.0805	0.6603	0.0066	141.2
8	whole	series	lång	rad	2	0.1546	0.2087	0.4516	0.0102	139.2
9	regulatory	framework	rättslig	ram	2	0.1168	0.1266	0.5619	0.0079	131.9
10	constructive	cooperation	god	samarbete	2	0.0470	0.0445	0.8323	0.0041	101.4
11	important	role	stor	roll	2	0.0933	0.0211	0.8691	0.0044	90.3
12	lead	committee	ansvarig	utskott	2	0.0236	0.1680	0.4987	0.0052	73.6
13	fellow	member	kär	kollega	2	0.2643	0.6567	0.1196	0.0182	62.8
14	absolute	priority	hög	prioritet	2	0.0737	0.1601	0.3575	0.0088	53.9
15	central	question	viktig	fråga	2	0.0149	0.1409	0.5068	0.0047	49.0
16	whole	range	lång	rad	2	0.1421	0.2087	0.1575	0.0102	44.6
17	last	year	gången	år	5	0.2675	0.2123	0.9221	0.0346	43.7
18	particular	case	konkret	fall	3	0.0583	0.0557	0.7535	0.0076	42.6
19	excellent	report	bra	betänkande	5	0.2209	0.0643	0.8447	0.0181	36.6
20	good	deal	hel	del	3	0.0266	0.2168	0.0371	0.0024	36.3
21	paramount	importance	stor	vikt	2	0.1651	0.1405	0.4416	0.0178	32.3
22	recent	year	gången	år	2	0.1575	0.2123	0.9221	0.0313	31.5
23	much	time	lång	tid	3	0.0306	0.2662	0.4821	0.0120	27.4
24	positive	result	god	resultat	2	0.0654	0.0616	0.6390	0.0102	24.9
25	less	time	kort	tid	2	0.0167	0.1730	0.4821	0.0078	22.7

3. Helping Learners with non-compositional items

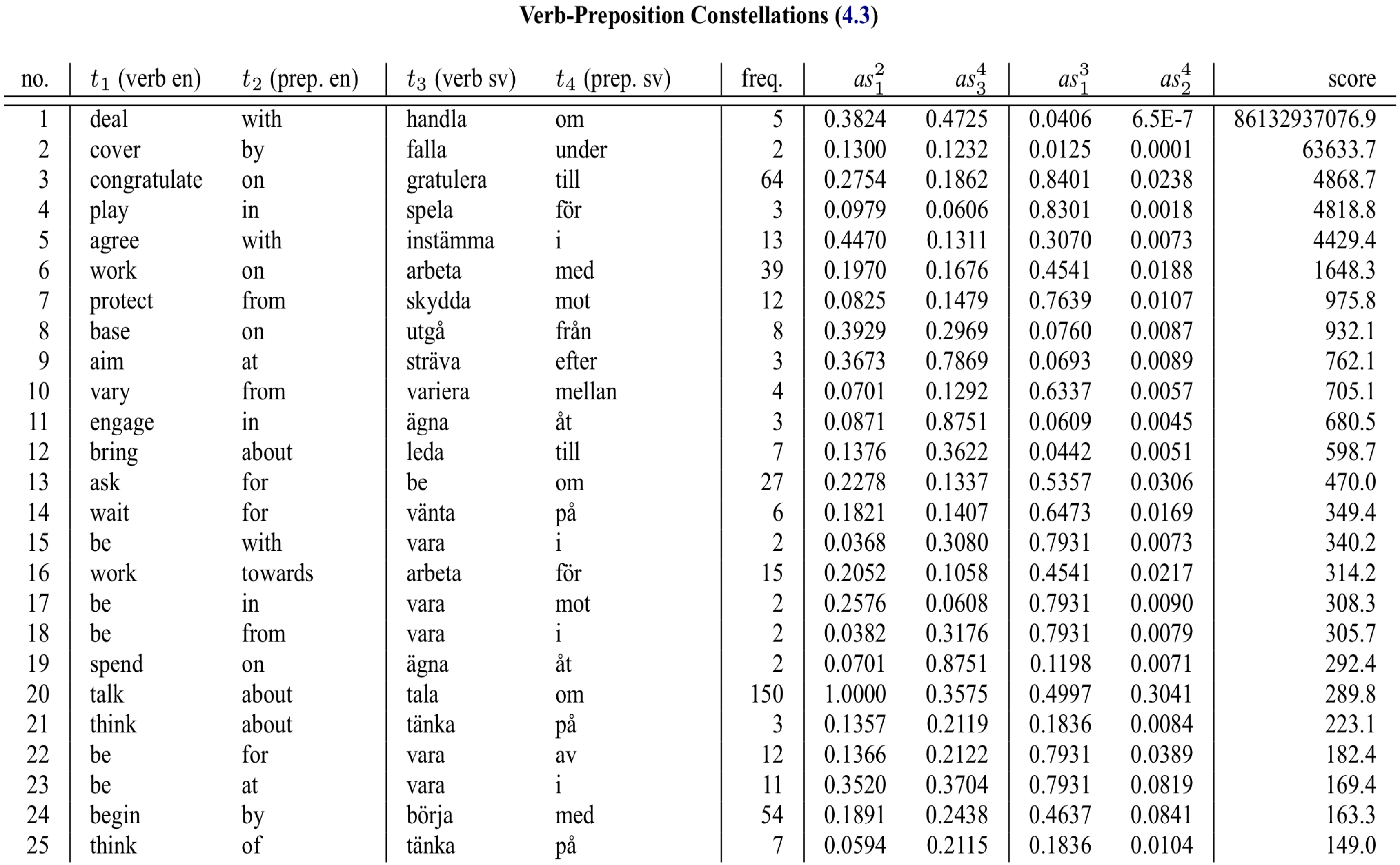
verb-object constellations: low strength on head=verb, high on dep



Verb-Object Constellations (4.2)

no.	t_1 (verb en)	t_2 (noun en)	t_3 (verb sv)	t_4 (noun sv)	freq.	as_1^2	as_3^4	as_1^3	as_2^4	score
1	have	responsibility	bära	ansvar	2	0.6526	0.9860	0.0021	0.6694	186877.9
2	have	question	ställa	fråga	4	0.3375	0.9768	0.0026	0.4664	90060.7
3	have	debate	föra	debatt	6	0.4452	0.3407	0.0032	0.6989	60222.9
4	play	role	ha	roll	5	1.0000	0.4895	0.0054	0.6997	58356.0
5	give	example	nämna	exempel	3	0.6751	0.6790	0.0052	0.6248	32362.0
6	have	result	ge	resultat	2	0.1987	0.6542	0.0025	0.5631	23064.3
7	give	example	ta	exempel	3	0.6751	0.2830	0.0044	0.6248	18835.1
8	have	discussion	föra	diskussion	2	0.4600	0.3453	0.0032	0.6033	18144.3
9	take	precedence	ha	företräde	3	0.7210	0.3544	0.0036	0.2892	17579.2
10	have	sympathy	känna	sympati	2	0.3260	0.5913	0.0037	0.5218	14688.4
11	do	damage	orsaka	skada	2	0.3518	0.9601	0.0054	0.5935	13678.4
12	lead	life	leva	liv	2	0.4910	0.9423	0.0068	0.6422	12867.2
13	achieve	solution	finna	lösning	2	0.2822	0.9910	0.0059	0.7118	11630.1
14	raise	issue	diskutera	fråga	3	0.9336	0.7690	0.0094	0.4719	11545.5
15	go	way	välja	väg	2	0.9071	0.7066	0.0063	0.2318	7501.2
16	fulfil	responsibility	ta	ansvar	2	0.3466	0.8815	0.0082	0.6694	6142.9
17	give	speech	hålla	tal	2	0.2396	0.6764	0.0047	0.3841	5741.6
18	hold	debate	ha	debatt	4	0.8837	0.2646	0.0109	0.6989	5551.5
19	accept	responsibility	ta	ansvar	15	0.5512	0.8815	0.0304	0.6694	5296.0
20	secure	majority	få	majoritet	2	0.7340	0.3258	0.0080	0.6983	5229.1
21	make	speech	hålla	tal	2	0.3974	0.6764	0.0063	0.3841	5178.7
22	bear	responsibility	ha	ansvar	2	0.7259	0.6271	0.0110	0.6694	5017.8
23	adopt	position	ta	ställning	4	0.8479	0.8543	0.0119	0.2393	4856.4
24	put	end	få	slut	6	0.9932	0.7120	0.0221	0.5538	4823.5
25	make	mistake	begå	misstag	5	0.8353	0.9947	0.0235	0.5856	4413.2

3. Helping Learners with non-compositional items



Verb-Preposition Constellations (4.3)

no.	t_1 (verb en)	t_2 (prep. en)	t_3 (verb sv)	t_4 (prep. sv)	freq.	as_1^2	as_3^4	as_1^3	as_2^4	score
1	deal	with	handla	om	5	0.3824	0.4725	0.0406	6.5E-7	86132937076.9
2	cover	by	falla	under	2	0.1300	0.1232	0.0125	0.0001	63633.7
3	congratulate	on	gratulera	till	64	0.2754	0.1862	0.8401	0.0238	4868.7
4	play	in	spela	för	3	0.0979	0.0606	0.8301	0.0018	4818.8
5	agree	with	instämma	i	13	0.4470	0.1311	0.3070	0.0073	4429.4
6	work	on	arbeta	med	39	0.1970	0.1676	0.4541	0.0188	1648.3
7	protect	from	skydda	mot	12	0.0825	0.1479	0.7639	0.0107	975.8
8	base	on	utgå	från	8	0.3929	0.2969	0.0760	0.0087	932.1
9	aim	at	sträva	efter	3	0.3673	0.7869	0.0693	0.0089	762.1
10	vary	from	variera	mellan	4	0.0701	0.1292	0.6337	0.0057	705.1
11	engage	in	ägna	åt	3	0.0871	0.8751	0.0609	0.0045	680.5
12	bring	about	leda	till	7	0.1376	0.3622	0.0442	0.0051	598.7
13	ask	for	be	om	27	0.2278	0.1337	0.5357	0.0306	470.0
14	wait	for	vänta	på	6	0.1821	0.1407	0.6473	0.0169	349.4
15	be	with	vara	i	2	0.0368	0.3080	0.7931	0.0073	340.2
16	work	towards	arbeta	för	15	0.2052	0.1058	0.4541	0.0217	314.2
17	be	in	vara	mot	2	0.2576	0.0608	0.7931	0.0090	308.3
18	be	from	vara	i	2	0.0382	0.3176	0.7931	0.0079	305.7
19	spend	on	ägna	åt	2	0.0701	0.8751	0.1198	0.0071	292.4
20	talk	about	tala	om	150	1.0000	0.3575	0.4997	0.3041	289.8
21	think	about	tänka	på	3	0.1357	0.2119	0.1836	0.0084	223.1
22	be	for	vara	av	12	0.1366	0.2122	0.7931	0.0389	182.4
23	be	at	vara	i	11	0.3520	0.3704	0.7931	0.0819	169.4
24	begin	by	börja	med	54	0.1891	0.2438	0.4637	0.0841	163.3
25	think	of	tänka	på	7	0.0594	0.2115	0.1836	0.0104	149.0

3. Helping Learners with non-compositional items

Try our Demo at

https://pub.cl.uzh.ch/projects/sparcling/constellations/

You can chose the collocation metric, and adapt the scoring function, e.g.

https://pub.cl.uzh.ch/projects/sparcling/constellations/dobj.php?dep_measure=t-score&al_measure=t-score&norm=tanhavg&score=as12*as34*as24/as13^2

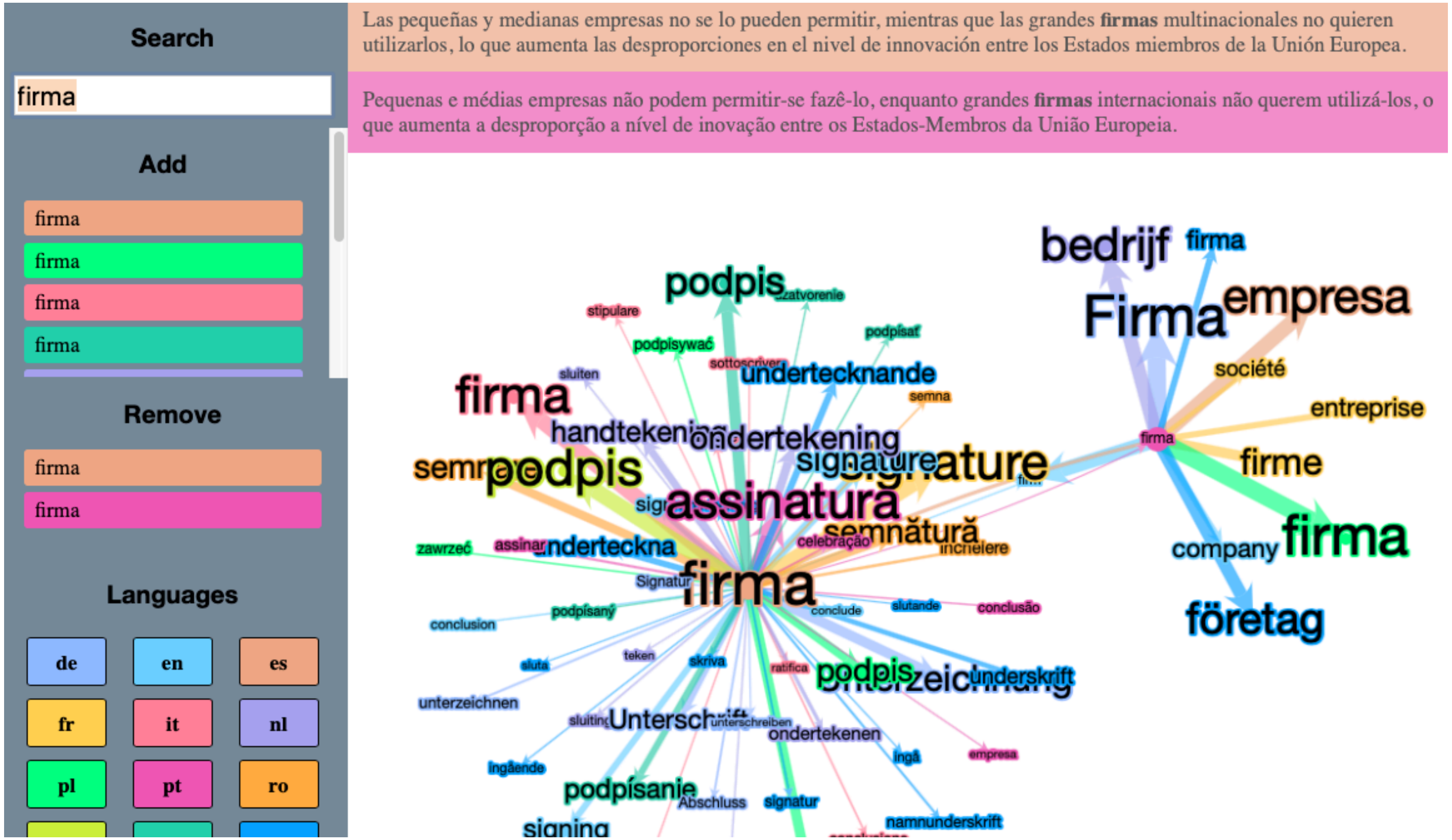
Our approach offers direct and indirect corpus use in combination:

- **Indirect corpus use**: Creating corpus-informed teaching materials, e.g. collocations dictionaries (Ackermann and Chen 2013; Durrant 2009; McGee 2012): students do not need to learn to use corpus interfaces, but contextualisation is limited.
- **Direct corpus use** improves learner competence in the area of collocations. Li (2017) concludes that "[t]his exposure to attested language data raises learners' awareness of using collocations in a more natural or near-native way" (p. 165)

4. Translational Scapes

False Friends are a frequent and difficult problem for language learners.
Most resources are in the form of dictionaries (Varela 2011) & incomplete

On the other hand, not all occurrences of “false friends” are incorrect. E.g.:
ES firma ↔ *PT firma* : demo at
https://pub.cl.uzh.ch/purl/alignment_overlap



Las pequeñas y medianas empresas no se lo pueden permitir, mientras que las grandes **firmas** multinacionales no quieren utilizarlos, lo que aumenta las desproporciones en el nivel de innovación entre los Estados miembros de la Unión Europea.

Pequenas e médias empresas não podem permitir-se fazê-lo, enquanto grandes **firmas** internacionais não querem utilizá-los, o que aumenta a desproporção a nível de inovação entre os Estados-Membros da União Europeia.

Search: firma

5. Outlook

- Overcome the data sparseness problem of surprisal by Deep Learning. Particulary BERT (Devlin 2018) shows promising results (~65%) on acceptability ratings (COLA, Warstadt 2018).
- Further integrate automatic parser (Schneider & Grigonyte 2018): Model fit of parser depends on learner level, but low predictive power
- Add such tools to existing writing systems: readbility, TTR, surprisal, specific grammatical warnings.
- Test the tools and resources on actual learners, in collaboration with didactics experts. Can the gap between implicit & explicit learning / direct & indirect corpus-use be closed?
- Playful approaches to learning, e.g.
    - cloze on the fly
    - predict sentence continuation
    - The Alternator

Q&A

Thank you for you attention!

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

TCC 2010

Reserved

The full Dozentenfoyer is between
13:00 – 15:00
for TCC 2010 reserved.

References

Ackermann, K. and Y. H. Chen (2013). “Devel- oping the Academic Collocation List (ACL): A corpus- driven and expert-judged approach.” In: *Journal of English for Academic Purposes* l2.4, pp. 235–247.

Ananiadou, Sophia, Kell, Douglas B., and Tsujii, Jun-ichi. 2006. “Text mining and its potential applications in systems biology”. *Trends in Biotechnology*, 24, 12, 571 – 579.

Aston, Guy and Burnard, Lou. 1998. The BNC Handbook. Exploring the British National Corpus with SARA. Edinburgh University Press, Edinburgh.

Bartsch, Sabine and Stefan Evert. 2014. “Towards a Firthian Notion of Collocation”. In A. Abel and L. Lemnitzer (eds.) *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern*. OPAL -- Online publizierte Arbeiten zur Linguistik (2/2014). Mannheim: Institut für Deutsche Sprache. 48-61.

Benson, M., Benson, E. & Ilson, R. 2009. *The BBI Combinatory Dictionary of English* (3rd ed.). Amsterdam: John Benjamins.

Choueka, Yaacov. 1988. "Looking for needles in a haystack". Proceedings of RIAO '88, 609-623.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805

Durrant, P. (2009). “Investigating the viability of a collocation list for students of English for aca- demic purposes”. In: *English for Specific Purposes* 28.3, pp. 157–169.

Ellis, Nick. 2012. Formulaic Language and Second Language Acquisition: Zipf and the Phrasal Teddy Bear. *Annual Review of Applied Linguistics*, 32, 17–44.

Erman, B. 2009. Formulaic language from a learner perspective: What the learner needs to know. In Corrigan, K., E. A. Moravcsik, H. Ouali, and K.M. Wheatley, eds. 2009. *Formulaic Language. Volume II: Acquisition, loss, psychological reality, and functional explanations*. Amsterdam/ Philadelphia: Benjamins. , 323–346.

Gardner, D. & Davies, M. 2007. “Pointing out frequent phrasal verbs: A corpus-based analysis”, *TESOL Quarterly: A Journal for Teachers of English to Speakers of Other Languages and of Standard English as a Second Dialect* 41(2), 339–359.

Glynn, Dylan. 2010. "Corpus-driven Cognitive Semantics. Introduction to the field". Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches, 1-42.

Gilquin, Gaëtanelle & Sylviane Granger. 2011. "From EFL to ESL: Evidence from the International Corpus of Learner English". Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap, 55-78.

Graën, Johannes, Dolores Batinic, and Martin Volk. 2014. “Cleaning the Europarl Corpus for Linguistic Applications”. In*: Proceedings of the Conference on Natural Language Processing (KONVENS).* Stiftung Universität Hildesheim, pp. 222–227.

Granger, Sylviane. 2009. Prefabricated patterns in advanced EFL writing: Collocations and formulae (OUP, 1998). In A. P. Cowie, editor, *Phraseology: Theory, analysis, and applications*. Kurosio Publishers, Tokyo, pages 185–204.

Granger, Sylviane, Dagneaux, Estelle, Meunier, Fanny, and Paquot, Magali. 2009. *International Corpus of Learner English v2 (Handbook+CD-Rom)*

References II

Ishikawa, S. 2009. Vocabulary in interlanguage: A study on corpus of English essays written by Asian university students (CEEAUS). In K. Yagi and T. Kanzaki, (eds): *Phraseology, corpus linguistics and lexicography: Papers from Phraseology 2009 in Japan.* Nishinomiya, Japan: Kwansei Gakuin University Press, 87–100.

Lehmann, Hans Martin & Gerold Schneider. 2011. “A large-scale investigation of verb-attached prepositional phrases”. In Hoffmann, S., Rayson, P. & Leech, G. (Eds.), *Studies in Variation, Contacts and Change in English, Volume 6: Methodological and Historical Dimensions of Corpus Linguistics*. Helsinki: Varieng. http://www.helsinki.fi/varieng/journal/volumes/06/

Levy, Roger and Jaeger, T. Florian. 2007. "Speakers optimize information density through syntactic reduction". Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems.

Li, S. (2017). “Using corpora to develop learn- ers’ collocational competence”. In: *Language Learning & Technology* 21.3, pp. 153–171.

McEnery, Tony, & Richard Xiao. 2011. What corpora can offer in language teaching and learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 364–380). London: Routledge.

McGee, I. (2012). “Collocation dictionaries as in- ductive learning resources in data-driven learn- ing: An analysis and evaluation”. In: *International Journal of Lexicography* 25.3, pp. 319– 361.

Millar, Neil. 2011. "The processing of malformed learner collocations". *Applied Linguistics*, 32, 2, 129-148.

Mukherjee, Joybrato and Sebastian Hoffmann. 2006. "Describing verb-complementational profiles of New Englishes: A pilot study of Indian English". English World-Wide, 27, 2, 147-173.

Nesselhauf, Nadja. 2009. “Co-selection phenomena across New Englishes: Parallels (and differences) to foreign learner varieties”. *English World-Wide* 30: 1-25.

Ng, Tou Hwee, Wu, Mei Siew, Wu, Yuanbin, Hadiwinoto, Christian, and Tetreault, Joel. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. 1-12.

Pawley, Andrew and Frances Hodgetts Syder. 1983. "Two Puzzles for Linguistic Theory: Native-like selection and native-like fluency". Language and Communication, 191-226.

Schneider, Gerold. 2008. Hybrid Long-Distance Functional Dependency Parsing. PhD Thesis, University of Zurich.

Schneider, Gerold and Grigonyte, Gintare. 2018. "From Lexical Bundles to Surprisal and Language Models: measuring the idiom principle on native and learner language". Applications of Pattern-driven Methods in Corpus Linguistics, 82, 15-55.

Schneider, Gerold and Marianne Hundt. 2009. “Using a parser as a heuristic tool for the description of New Englishes.” In *Proceedings of Corpus Linguistics 2009*, Liverpool.

Schneider, Gerold and Lena Zipp. 2013. “Discovering new verb-preposition combinations in New Englishes”. *Studies in Variation, Contacts and Change in English* 13. Available at http://www.helsinki.fi/varieng/series/volumes/13/schneider_zipp.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. John Benjams, Amsterdam.

Warstadt, Alex, Singh, Amanpreet, and Bowman, Samuel R. 2018. "Neural Network Acceptability Judgments". arXiv preprint arXiv:1805.12471.