Les collocations statistiques au service de la recherche en acquisition des langues étrangères

Magali Paquot (CECL) & Hubert Naets (CENTAL)

Phraseology

- Language is essentially made up of word combinations that constitute single or preferred choices
  - *raise + issue, carry + implication*
  - *fill + in, make + out*
  - *cut from whole cloth, cut it close*
  - *It has been suggested that …, as exemplified by*
- Word combinations play crucial roles in language acquisition, proficiency & fluency

2

Sinclair (1991), Ellis (1996), Biber et al. (1999), Wray (2002), Stefanowitsch & Gries (2003), Schmitt (2004), Goldberg (2006), Granger & Paquot (2008), Ellis & Cadierno (2009), Römer (2009), Bybee & Beckner (2012)

Foreign language learning

- Phraseological units remain a source of errors even at advanced proficiency levels
  - (verb-object) collocations and phrasal verbs

- Higher proficiency is usually characterized by:
  - A higher rate of use of native-like collocations
  - A lower rate of use of repeated sequences

Paquot & Granger (2012), Ellis et al (2015), Oksefjell Ebeling & Hasselgård (2015)

3

Statistical collocations

- "co-occur more often than their respective frequencies and the length of text in which they appear would predict" (Jones & Sinclair, 1974: 19)

	$V = v$	$V \neq v$	
$U = u$	O_{11}	O_{12}	$= R_1$
$U \neq u$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

observed frequencies

	$V = v$	$V \neq v$
$U = u$	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$
$U \neq u$	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$

expected frequencies

4

'get': statistical collocations (PMI)

$$\textbf{MI} = \log \frac{O_{11}}{E_{11}}$$

Collocation parameters:			
Information:	collocations	Statistics:	Mutual information
Collocation window span:	1 Right - 1 Right	Basis:	spoken texts only
Freq(node, collocate) at least:	5	Freq(collocate) at least:	5
Filter results by:	Specific collocate:	and/or tag: no restrictions	Submit changed parameters Go!

There are 35634 different types in your collocation database for "[lemma="(get)_VERB"%c]". (Your query "{get/V}" in written texts returned 117885 hits in 2769 different texts)

No.	Word	Total No. in spoken texts	Expected collocate frequency	Observed collocate frequency	In No. of texts	Mutual information value
1	underway	24	0.236	136	93	9.1705
2	fatter	5	0.049	19	17	8.594
3	impatient	12	0.118	30	29	7.9899
4	undressed	13	0.128	28	26	7.7749
5	acquainted	9	0.089	19	19	7.746
6	bogged	16	0.157	32	31	7.668
7	rid	755	7.425	1337	713	7.4925
8	yer	10	0.098	17	13	7.4336
9	stung	5	0.049	8	7	7.3461
10	airborne	12	0.118	19	13	7.331

5

'get': statistical collocations (t-score)

$$\textbf{t-score} = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$$

Collocation parameters:			
Information:	collocations	Statistics:	T-score
Collocation window span:	1 Right - 1 Right	Basis:	spoken texts only
Freq(node, collocate) at least:	5	Freq(collocate) at least:	5
Filter results by:	Specific collocate:	and/or tag: no restrictions	Submit changed parameters Go!

There are 35634 different types in your collocation database for "[lemma="(get)_VERB"%c]". (Your query "{get/V}" in written texts returned 117885 hits in 2769 different texts)

No.	Word	Total No. in spoken texts	Expected collocate frequency	Observed collocate frequency	In No. of texts	T-score value
1	a	206,201	2,027.763	10514	1794	82.762
2	to	233,691	2,298.098	8642	1585	68.2416
3	out	29,679	291.861	3817	1149	57.0578
4	the	409,714	4,029.093	9264	1842	54.3888
5	into	11,007	108.242	2786	1101	50.7319
6	back	15,863	155.995	2513	880	47.018
7	up	34,929	343.489	2786	896	46.275
8	on	81,082	797.354	3262	1105	43.1532
9	away	5,634	55.404	1900	862	42.3179
10	rid	755	7.425	1337	713	36.362
11	his	14,046	138.127	1304	664	32.2859
12	some	20,589	202.471	1349	688	31.2161
13	through	7,967	78.347	1014	612	29.383

6

Statistical collocations in foreign language learning research

Learner corpus	MI	BNC	MI
new nation	?	new nation	2.11
a great	?	a great	3.88
attractive reading	?	attractive reading	/
there are	?	there are	4.94
we can	?	we can	4.36
economic point	?	economic point	0.99
fact that	?	fact that	5.16
hand there	?	hand there	0.34
is obvious	?	is obvious	2.91
is probable	?	is probable	4.62
possibility to	?	possibility to	-1.57
the unification	?	the unification	1.52
we really	?	we really	2.15

7

Durrant & Schmitt (2009)

- Compared to native speakers, learners
  - overuse collocations identified by high t-scores
    - *good example, long way, hard work*
  - underuse collocations identified by high PMI scores
    - *densely populated, bated breath, preconceived notions*

8

Granger & Bestgen (2014)

- Learner corpus: *International Corpus of Learner English* (ICLE, Granger et al., 2009)

- Compared to intermediate learners, advanced EFL learners have
  - a lower proportion of collocations identified by high t-scores
    - High-frequency, simple, large collocational network
  - a higher proportion of collocations identified by high PMI scores
    - Low frequency, more sophisticated, collocational restrictions

9

From a « positional » model …

- **Adjacent** premodifier-noun word pairs (i.e. adj-noun and noun-noun combinations) (Siyanova & Schmitt, 2008; Durrant & Schmitt, 2009)
- Bigrams (i.e. **contiguous** pairs of words) (Bestgen & Granger, 2014; Granger & Bestgen, 2014)
  - *Yesterday they won the Spanish lottery*
    - *yesterday + they, they + won, won + the, the + Spanish, Spanish+ lottery*

- V + Obj, S+ Verb, V + Particle, Adv. + V, …

10

… to a « relational » model of statistical collocations

- Co-occurring words appear in a specific structural relation (Evert, 2005)

- *Yesterday they won the Spanish lottery.*
  - ~~*Yesterday they, won the, the Spanish*~~
  - *won + lottery*

11

METHOD

12

Corpus processing

Ref. corpus + learner corpus	0. Corpus cleaning	• In-house Perl programs
	1. Lemmatisation and part-of-speech tagging	• Stanford CoreNLP; TreeTagger
	2. Parsing and extraction of dependencies	• Stanford CoreNLP; MaltParser
	3. Simplification of POS tags, computing frequencies, etc.	• In-house Perl programs
	4. Data storing	• Redis
Ref. corpus	5. Calculation of association measures between a pair of words in a particular Stanford typed dependency	• Ngram Statistics Package (NSP) • In-house Perl programs

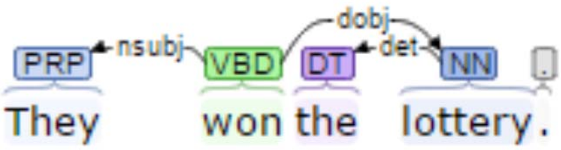
13

1. Lemmatisation and part-of-speech tagging

- *They won the lottery*.
- *They*[they.PRP] *won*[win.VBD] *the*[the.DT] *lottery*[lottery.NN].

14

2. Parsing and extraction of dependencies



- nsubj(won,they)
- dobj(won,lottery)
- det(lottery,the)

De Marneffe & Manning (2010)

15

3. In-house Perl programs

***They*[they.PRP] *won*[win.VBD] *the*[the.DT] *lottery*[lottery.NN]**	**nsubj(*won,they*) dobj(*won,lottery*) det(*lottery,the*)**
Lemma + simplified POS	Dependencies + frequencies

nsubj(win.VB,they.PRP) 8
dobj(win.VB,lottery.NN) 4
det(lottery.NN,the.DT) 25

16

4. Association scores

- Assign to each word combination (type) extracted from the learner corpus under study an association score computed on the basis of a reference corpus
  - Pointwise mutual information
    - Freq > 4 in reference corpus
- Compute mean PMI scores for each dependency relations in each learner text (cf. Bestgen & Granger, 2014)

17

STUDY 1: PAQUOT (SUBMITTED)

18

RQs

- To what extent can measures of phraseological sophistication (i.e. statistical collocations as identified by MI scores) be used to describe L2 performance at different proficiency levels?
  - amod, advmod, dobj
- How do measures of phraseological sophistication compare with measures of lexical sophistication?

19

VESPA-FR-LING

- *Varieties of English for Specific Purposes Database* (VESPA)
- http://www.uclouvain.be/en-cecl-vespa.html

Per institutional level	Number of files	Total number of words	Means
B2	25	86,472	3,588
C1	62	216,283	3,488
C2	11	33,994	3,090
Total	**98**	**336,749**	**3,436**

20

L2 research corpus (L2RC)

- 16 major journals in L2 research (1980-2014)
  - Applied Linguistics, Applied Language Learning, Applied Psycholinguistics, Bilingualism: Language and Cognition, The Canadian Modern Language Review, Foreign Language Annals, Journal of Second Language Writing, Language Awareness, Language Learning, Language Learning and Technology, Language Teaching Research, The Modern Language Journal, Second Language Research, Studies in Second Language Acquisition, System, TESOL Quarterly
- 7,765 texts
- 66,218,913 words (363 Mio)
- 49,754,608 dependencies

21

Measures of lexical sophistication

	Lexical sophistication	Formula
LS1	Lexical sophistication-I	N_{slex}/N_{lex}
LS2	Lexical sophistication-II	T_s/T
VS1	Verb sophistication	T_{sverb}/N_{verb}
CVS1	Corrected VSI	$T_{sverb}/\sqrt{N_{verb}}$
VS2	Verb sophistication-II	T^2_{sverb}/N_{verb}

Lexical Complexity Analyzer (Lu, 2012)

22

Learner group comparisons

- Shapiro-Wilk normality tests
  - ANOVAs + Tukey contrasts
  - Kruskal-Wallis rank sum tests
- p < 0.05 (with Bonferroni corrections to correct for multiple comparisons)

23

amod: adjectival modifier

- « Sam eats red meat. » → amod(meat,red)

  NN JJ

F(2,98)= 5,642, p = 0.00484, eta squared (η^2)= 0,1061

	N	Mean PMI	sd
B2	25	2.42	0.33
C1	62	2.62	0.42
C2	11	2.9	0.44

B2 – C1

C1 – C2

B2 – C2 **

24

Examples of *amod* dependencies

- pmi > 6 : *overwhelming majority, hasty conclusion, integral part, slight predominance, keen interest, exhaustive list, wide range, illustrative example, chronological order, wide variety, spontaneous speech, next section, possible explanation, large majority, significant difference, clear preference*
- pmi = 1: *main function, only conclusion, final part, common history, different field, same number, enough material, theoretical definition, common word, long word, real power, specific form, common method, certain way, different function, general definition, simple form*

25

advmod: adverbial modifier

- advmod(unprecedented+JJ,totally+RB)
- advmod(enough+RB,strangely+RB)
- advmod(root+VB,firmly+RB)

F(2,98)= 6.382 , p = 0.00251, eta squared (η^2)= 0,1184

	N	Mean PMI	sd
B2	25	1.18	0.30
C1	62	1.39	0.28
C2	11	1.48	0.20

B2 – C1 **

C1 – C2

B2 – C2 **

26

Examples of *advmod* dependencies

- pmi > 7 :
  - advmod(incorrect+JJ,grammatically+RB), advmod(significant+JJ,statistically+RB), advmod(rightly+RB,quite+RB), advmod(understandable+JJ,perfectly+RB), advmod(distribute+VB,evenly+RB), advmod(evolve+VB,constantly+RB)
- pmi = 1:
  - advmod(interesting+JJ,quite+RB), advmod(possible+JJ,also+RB), advmod(puzzling+JJ,more+RB)

27

dobj: direct object

- dobj(make+VB,statement+NN)

F(2,98)= 8.636, p = 0.000358, eta squared (η^2)= 0,1538

	N	Mean PMI	sd
B2	25	1.79	0.39
C1	62	1.97	0.40
C2	11	2.38	0.36

B2 – C1
C1 – C2 **
B2 – C2 **

28

Examples of *dobj* dependencies

- pmi > 7:
  - dobj(arouse+VB,curiosity+NN), dobj(fill+VB,gap+NN), dobj(serve+VB,purpose+NN), dobj(pay+VB,attention+NN), dobj(play+VB,role+NN), dobj(divert+VB,attention+NN), dobj(corroborate+VB,finding+NN), dobj(avoid+VB,misunderstand+NN)
- Pmi = 1:
  - dobj(have+VB,function+NN), dobj(consider+VB,characteristic+NN), dobj(have+VB,characteristic+NN), dobj(classify+VB,adjective+NN), dobj(mention+VB,agent+NN)

29

Lexical sophistication across proficiency levels

	B2		C1		C2		Between-group comparisons
	Mean	SD	Mean	SD	Mean	SD	
LS1	0.43	0.04	0.42	0.05	0.43	0.05	F(2,98)=0.10, p = 0.91
LS2	0.35	0.04	0.34	0.05	0.37	0.02	F(2,98)=1.98, p = 0.14
VS1	0.09	0.02	0.09	0.03	0.11	0.03	H(2,98)=5.64, p = 0.06
CVS1	1.27	0.33	1.26	0.36	1.43	0.30	F(2,98)=1.21, p = 0.30
VS2	3.43	1.84	3.41	1.98	4.28	1.67	H(2,98)=3.24, p = 0.20

30

Interim summary

- Mean PMIs: B2 > C1 > C2
  - *amod*
    - B2 / C2
  - *advmod*
    - Intermediate vs. advanced: B2 / C1-C2
  - *dobj*
    - B2 – C1 / C2
- Lexical sophistication: no linear increase

31

STUDY 2: PAQUOT & NAETS (2015)

32

Objective

- Investigate whether statistical collocations can be used to trace phraseological development in a longitudinal learner corpus

33

UCL component of LONGDALE (Meunier & Littré, 2013)

- Undergraduate students of English in Louvain
- French-speaking learners
- Argumentative essays (8 topics)
- Oxford Quick Placement Test > CEFR

	Number of texts (with OQPTs)
Year 1	184
Year 2	109
Year 3	124
Total	**417**

34

Mixed-effects modeling

- « Mixed effects models are robust against missing data » (Cunnings & Finlayson, 2015: 162)
- Assess the influence of fixed effects (= ***time, proficiency, topic***), while taking into account any random variation observed (= ***random variance across the participants tested***).

35

Mixed-effects modelling: technical details

- R (R Core Team, 2014) + ggplot2, lme4, lmerTest, effects, MuMIn packages
- Model selection procedure (Zuur et al, 2009; Gries, 2015b:112):
  - begin with a model that contains the most comprehensive fixed effects structure that can be fit given the variables to be explored and find the optimal random-effects structure (varying intercepts for one or more predictors and/or varying slopes for one or more predictors); and,
  - once the optimal random-effects structure has been found, find the optimal fixed-effects structure.

36

Reference corpus: ENCOW14 (AX version, Schäfer, 2015)

- Web corpus
  - 9,578,828,861 tokens; 425,374,806 sentences
  - Stanford typed dependencies: Malt Parser

↓

- LONGDALE
  - POS-tagged and lemmatized with Tree Tagger
  - Parsed with Malt Parser

37

dobj dependencies: pmi values

- High PMI scores (>= 7)
  - commit + crime, ride + horse, watch + television, browse + web, cure + disease, park + car, solve + problem, earn + living, attend + meeting, mow + lawn, draw + conclusion, serve + purpose, seek + refuge, raise + awareness
- Low PMI scores (<= 2)
  - design + society, imagine + phenomenon, win + conflict, develop + science, suggest + idea, dream + life, find + place, have + dream, have + time, have + friend, have + power, buy + thing, buy + anything, desire + something, want + money

38

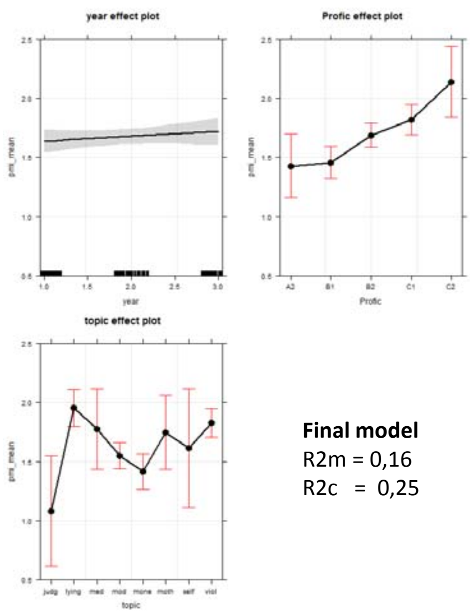
dobj: final mixed-effect model (av. pmi values)

model.final <- lmer(pmi_mean ~ year + Profic + topic + (1|task_partid), data=LCR2015_dobj)

```
Random effects:
 Groups      Name        Variance Std.Dev.
 task_partid (Intercept) 0.0386   0.1965
 Residual                0.3509   0.5923
Number of obs: 417, groups:  task_partid, 237

Fixed effects:
            Estimate Std. Error        df t value Pr(>|t|)
(Intercept)  0.76327    0.27721 403.80000   2.753 0.006164 **
year         0.04145    0.04336 350.10000   0.956 0.339764
ProficB1     0.02899    0.14716 402.60000   0.197 0.843913
ProficB2     0.26041    0.14734 403.90000   1.767 0.077908 .
ProficC1     0.39189    0.15533 390.10000   2.523 0.012035 *
ProficC2     0.71188    0.20744 389.50000   3.432 0.000664 ***
topiclying   0.87267    0.24831 397.20000   3.514 0.000491 ***
topicmed     0.69102    0.29146 400.40000   2.371 0.018219 *
topicmod     0.46677    0.24423 401.40000   1.911 0.056694 .
topicmone    0.33572    0.24960 404.00000   1.345 0.179369
topicmoth    0.66320    0.28616 400.30000   2.318 0.020976 *
topicself    0.53147    0.34626 399.70000   1.535 0.125597
topicviol    0.74659    0.24585 403.90000   3.037 0.002546 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

39



Final model
R2m = 0,16
R2c = 0,25

40

Fixed effects on av. PMI values in dobj dependencies

Work in progress!

- Phraseological indices
  - Means per text – crude measures
    - SD
    - Collocational bands (%)
  - Other association measures
- Statistical analyses
  - Model to assess the respective effects of the different measures
- Reference corpus
  - Compare results based on BNC, ENCOW14, L2RC
- L2 Dutch, L2 French

41

IMPLICATIONS

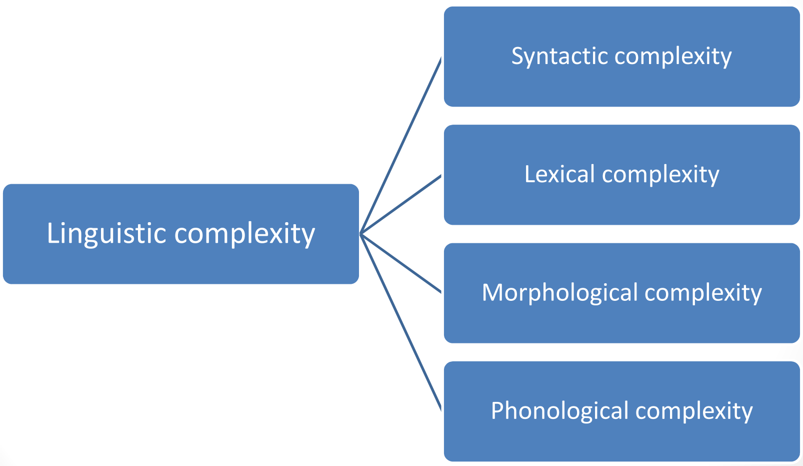
42

Phraseology

- Essential dimension of L2 writing quality (and probably even more so at the more advanced proficiency levels)
- Influence overall perceptions of language proficiency by expert evaluators
  - Adjacent proficiency levels

43

Phraseology: missing dimension in linguistic complexity



Bulté & Housen (2012)

44

Phraseological complexity (Paquot, submitted)

- I'll **meet** you in the bar later.
- I **met** *up with* John as I left the building.
- This app has different versions to **meet** different *needs*.
- To **meet** customer *expectations*, several initiatives have been taken.
- If you **meet** your *target*, congratulate yourself.
- 'Here I believe my brother has **met** *his Waterloo*,' she murmured.
- There is *more than* **meets** *the eye*.
- Many students are finding it difficult *to make ends* **meet**.
- *Nice to* **meet** *you*!
- *It's a pleasure to* **meet** *you*!

45

Language teaching & testing

- Not a single mention of 'collocations', 'phraseology', or 'formulaic sequences' in the *Structured Overview of all CEFR scales* published by the Council of Europe (2001)
- Phraseological complexity should feature more prominently in language proficiency descriptors and second language assessment rubrics than it currently does.

46

Automated assessment

- Phraseological indices (based on collocations, ngrams, collostructions, etc.) could be used to augment the set of linguistic indices used to automatically score L2 productions
  - e-rater® (ETS):
    - No assessment of the variability, sophistication, etc. of word combinations
- Context-sensitive measures
  - Mode
  - Genre
  - Topic

47

Thank you very much!

Questions? Comments? Suggestions?

References

- Paquot, M. (submitted). Phraseological competence: a missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. Special issue of Language Assessment Quarterly on 'Language tests for academic enrolment and the CEFR' (guest editors: Bart Deygers, Cecilie Hamnes Carlsen, Nick Saville & Koen Van Gorp). Submitted (invitation)
- Paquot, M. (submitted). The lexis-grammar interface in interlanguage complexity research. Special issue of Second Language Research on 'Multiple approaches to L2 Complexity' (guest editors: Alex Housen and Bastien De Clercq). Submitted (invitation)
- Paquot, M. & Naets, H. (2015a). Using relational co-occurrences to trace phraseological development in a longitudinal corpus. Paper presented at the 25th EUROSLA conference, 27-29 August 2015, Aix-en-Provence, France.
- Paquot, M. & Naets, H. (2015b). Adopting a relational model of co-occurrences to trace phraseological development. Paper presented at the 3rd Learner Corpus Research Conference, 11-13 September 2015, The Netherlands.