Phraseological complexity in EFL learner writing across proficiency levels

Magali Paquot (FNRS – UCLouvain)

Introduction

- Language is essentially made up of word combinations that constitute single choices and words acquire meanings from their context (Sinclair, 1991; Biber et al., 1999; Wray, 2002)

- Word combinations play crucial roles in language acquisition, processing, fluency, idiomaticity and change (e.g. Ellis, 1996; Sinclair, 1991; Wray, 2002; Stefanowitsch & Gries, 2003; Schmitt, 2004; Goldberg, 2006; Ellis & Cadierno, 2009; Römer, 2009; Bybee & Beckner, 2012).

L2 complexity research

- Largely impervious to these theoretical and empirical developments.
- L2 complexity is admittedly no longer narrowed down to syntactic complexity (e.g. Bulté & Housen, 2012)
  - Phonology, lexis, morphology
- No systematic attempt to theorize and operationalize linguistic complexity at the level of word combinations
- Unfortunate as complexity = "one of the major research variables in applied linguistic research" (Housen & Kuiken, 2009)

- 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*!

Research programme

- Define and circumscribe the linguistic construct of phraseological complexity

- Theoretically and empirically demonstrate its relevance for second language theory in general and L2 complexity research in particular

Dimensions of complexity

- DIVERSITY
  - Breadth of knowledge
  - How many words or structures are known
  - Number of unique words in a text (e.g. TTR, D)
  - Absolute complexity
- SOPHISTICATION
  - Depth of knowledge
  - How elaborate or difficult the words and structures are
  - Frequency bands
  - Relative complexity

Bulté & Housen (2012), Ortega (2012), Wolfe-Quintero et al (1998)

Phraseological complexity

- Variety/diversity and sophistication
  - A learner text with a wide range of (target-like) phraseological units and a high proportion of relatively unusual or sophisticated units will be said to be more complex than one where the same few basic word combinations are often repeated.

- Working definition
  - The range of phraseological units that surface in language production and the degree of sophistication of such forms (cf. Ortega, 2003)

Paquot (2017)

- RQ1: To what extent can measures of phraseological complexity be used to describe L2 performance at different proficiency levels?

- RQ2: How do measures of phraseological complexity compare with traditional measures of syntactic and lexical complexity?

DATA AND METHODOLOGY

'Advancedness' in academic settings

- *Varieties of English for Specific Purposes Database* (VESPA)
  - <u>L1s</u>: Dutch, French, German, Italian, Norwegian, Spanish, Swedish
  - <u>Disciplines</u>: linguistics, business, engineering, …
  - <u>Genres</u>: research papers, reports
  - <u>Levels</u>: BA + MA

http://www.uclouvain.be/en-cecl-vespa.html

VESPA-FR-LING

Per proficiency 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**

https://uclouvain.be/en/research-institutes/ilc/cecl/vespa.html

Phraseological complexity

- Word combinations used in three types of grammatical dependency

amod	Adjectival modifier	*She has black hair* amod(hair+NN,black+JJ)
advmod	Adverbial modifier	*She has very black hair* advmod(black+JJ,very+RB) *Repeat less quickly.* advmod(quickly+RB,less+RB) *She eats slowly.* advmod(eat+VBZ,slowly+RB)
dobj	Direct object	*He won the lottery.* dobj(win+VV,lottery+NN)

Corpus workflow

1. Lemmatisation and part-of-speech tagging	Stanford CoreNLP: a suite of core NLP tools
2. Parsing and extraction of dependencies	
3. Simplification of POS tags, computing frequencies, etc.	In-house Perl programs

13

Phraseological diversity

	Phraseological diversity	Formula
amod_RTTR	Root TTR for amod dependencies	$T_{amod}/\sqrt{N_{amod}}$
advmod_RTTR	Root TTR for advmod dependencies	$T_{advmod}/\sqrt{N_{advmod}}$
dobj_RTTR	Root TTR for dobj dependencies	$T_{dobj}/\sqrt{N_{dobj}}$

Phraseological sophistication

- "selection of low-frequency [word combinations] that are appropriate to the topic and style of writing, rather than just general, everyday vocabulary", which "includes the use of technical terms (…) as well as the kind of uncommon [word combinations] that allow writers to express their meanings in a precise and sophisticated manner" (Read, 2000: 200).

- No general list of word combinations and their frequencies in English.

Phraseological sophistication I: Academic collocations

- The *Academic Collocation List* (Ackermann & Chen, 2013)
    - written curricular component of the Pearson International Corpus of Academic English (PICAE, over 25 million words)
    - the 2,469 most frequent and (according to its authors) pedagogically relevant cross-disciplinary lexical collocations in written academic English
    - http://pearsonpte.com/research/academic-collocation-list/

Phraseological sophistication I

	Phraseological sophistication	Formula
LS1amod	Lexical sophistication-I (amod)	Namods/ Namod
LS1advmod	Lexical sophistication-I (advmod)	Nadvmods/Nadvmod
LS1dobj	Lexical sophistication-I (dobj)	Ndobjs/Ndobj
LS2amod	Lexical sophistication-II (amod)	Tamods/ Tamod
LS2advmod	Lexical sophistication-II (advmod)	Tadvmods/Tadvmod
LS2dobj	Lexical sophistication-II (dobj)	Tdobjs/Tdobj

Phraseological sophistication II: MI scores

- Average pointwise mutual information (MI) score for amod, advmod and dobj dependencies.
  - compares the probability of observing word a and word b together with the probabilities of observing a and b independently (Church and Hanks 1990).
- Phraseological units that score very high on this measure have quite distinctive meanings (cf. Ellis et al., 2008)
  - *citric acid cycle, come into play, that leads to*
- Native speakers have been shown to be "attuned to these constructions as packaged wholes" (ibid).

Statistical collocations in SLA

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

Siyanova & Schmitt (2008), Durrant & Schmitt (2009), Groom (2009), Bestgen & Granger (2014), Granger & Bestgen (2014)

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*

Granger & Bestgen (2014)

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

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

- *bad weather, cold weather*

- *severe weather, extreme weather, stormy weather, windy weather* and *wintry weather*

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

Thanks to Luke Plonsky from Northern Arizona University for sharing the L2RC!

Corpus processing workflow

	Tools	Corpus
1. Lemmatisation	Stanford CoreNLP	L2RC + VESPA
2. Part-of-speech tagging		
3. Parsing		
4. Extraction of dependencies		
5. Simplify POS tags	In-house Perl programs	L2RC + VESPA
6. Compute corpus-based frequencies		
7. Compute MI scores for each pair of words in a dependency	Ngram Statistics Package	L2RC
8. Assign MI scores computed on the basis of the L2RC to each pair of words in a dependency in each learner text	In-house Perl program	VESPA
9. Compute mean MI scores for each learner text	R	VESPA

Thanks to Hubert Naets (CENTAL, UCLouvain) for his invaluable help!

Phraseological sophistication II

	Phraseological sophistication	Formula
mMIamod	Mean MI score for amod dependencies	Σ MIamod / Namod
mMIadvmod	Mean MI score for advmod dependencies	Σ MIadvmod / Nadvmod
mMIobj	Mean MI score for dobj dependencies	Σ MIdobj / Ndobj

Syntactic complexity

	Syntactic complexity (sophistication)
C/T	Clauses per T-unit
DC/T	Dependent clauses per T-unit
DC/C	Dependent clauses per clause
MLC	Mean length of clause
VP/T	Verb phrases per T-unit
CN/T	Complex nominals per T-unit
CN/C	Complex nominals per clause

- *L2 Syntactic Complexity Analyzer* (Lu, 2010)

25

Lexical diversity

	Lexical diversity	Formula
RTTR	Root TTR	$T/\sqrt{N}$
LV	Lexical word variation	Tlex/Nlex
CVV1	Corrected VV1	$Tverb/\sqrt{2Nverb}$
VV2	Verb variation-II	Tverb/Nlex
NV	Noun variation	Tnoun/Nlex
AdjV	Adjective variation	Tadj/Nlex
AdvV	Adverb variation	Tadv/Nlex

- *Lexical Complexity Analyzer* (Lu, 2012)

Lexical sophistication

	Lexical sophistication	Formula
LS1	Lexical sophistication-I	Nslex/Nlex
LS2	Lexical sophistication-II	Ts/T
VS1	Verb sophistication	Tsverb/Nverb
CVS1	Corrected VSI	Tsverb/√Nverb
VS2	Verb sophistication-II	T^2sverb/Nverb

- *Lexical Complexity Analyzer* (Lu, 2012)

RESULTS & DISCUSSION

Phraseological diversity

	B2		C1		C2		Between-group comparisons
	Mean	SD	Mean	SD	Mean	SD	
amod_RTTR	10.56	2.40	10.30	2.33	11.09	1.84	$F_{(2,98)}=0.66$, p = 0.52
advmod_RTTR	11.23	1.70	11.55	2.14	11.49	1.56	$F_{(2,98)}=0.09$, p = 0.95
dobj_RTTR	9.62	1.78	9.02	1.59	8.75	1.51	$H_{(2,98)}=1.61$, p = 0.21

Alpha set at 0.05/3 = 0.017

- No statistically significant difference

Phraseological sophistication I

	B2		C1		C2		Between-group comparisons
	Mean	SD	Mean	SD	Mean	SD	
LS1amod	0.03	0.02	0.03	0.03	0.04	0.02	H(2,98)=4.25, p = 0.12
LS1advmod	0.003	0.004	0.007	0.01	0.01	0.02	H(2,98)=4, p = 0.14
LS1dobj	0.009	0.01	0.009	0.01	0.02	0.02	H(2,98)=5.09, p = 0.08
LS2amod	0.03	0.02	0.03	0.02	0.04	0.02	H(2,98)=3.06, p = 0.22
LS2advmod	0.004	0.005	0.006	0.007	0.01	0.01	H(2,98)=3.55, p = 0.17
LS2dobj	0.007	0.007	0.009	0.009	0.01	0.01	H(2,98)=4.95, p = 0.08

Alpha set at 0.05/6 = 0.008

- (Linear) increase
- No statistically significant difference

30

Phraseological sophistication II

	amod		advmod		dobj	
	Mean MI	SD	Mean MI	SD	Mean MI	SD
B2	2.42	0.33	1.18	0.30	1.79	0.39
C1	2.62	0.42	1.39	0.28	1.97	0.40
C2	2.9	0.44	1.48	0.20	2.38	0.36

High vs. low MI scores

- amod dependencies with MI > 3: *overwhelming majority, hasty conclusion, integral part, slight predominance, keen interest, exhaustive list, wide range, illustrative example, chronological order*

- amod dependencies with MI = 1: *main function, only conclusion, final part, common history, different field, same number, enough material, theoretical definition, common word, long word*

- advmod dependencies with MI > 3: *grammatically incorrect, statistically significant, quite rightly, perfectly understandable, evenly + distribute, constantly + evolve*

- advmod dependencies with MI = 1: *quite interesting, also possible, more puzzling*

- dobj dependencies with MI > 3: *arouse + curiosity, fill + gap, serve + purpose, pay + attention, play + role, divert + attention, corroborate + finding, avoid + misunderstand*

- dobj dependencies with MI = 1: *have + function, consider + characteristic, have + characteristic*

amod dependencies

$F(2,98) = 5{,}642$, $p = 0{,}00484$, eta squared $= 0{,}1062$

	Estimate	Std. Error	t value	Pr(>|t|)
C1 – B2	0.20	0.10	2.059	0.10067
C2 – B2	0.48	0.15	3.308	0.00372 **
C2 – C1	0.28	0.13	2.168	0.07914

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05
(Adjusted p values reported -- single-step method)

advmod dependencies

$F(2,98) = 6{,}382$, $p = 0{,}00251$ eta squared $= 0{,}1184$

	Estimate	Std. Error	t value	Pr(>|t|)
C1 − B2	0.21	0.07	3.126	0.00641 **
C2 − B2	0.30	0.10	2.989	0.00946 **
C2 − C1	0.10	0.09	1.042	0.54530

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05
(Adjusted p values reported -- single-step method)

Examples of *advmod* dependencies with MI score > 6

- *mutually exclusive, fiercely debated, scarcely tenable, evenly distributed, firmly rooted, deeply rooted, stylistically heavy, regret profoundly, intimately intertwined, defined unclearly, disproportionately large, strangely enough, totally unprecedented, seriously endangered, officially approved, roughly equivalent, almost exclusively, rely heavily, vary enormously, statistically significant, linguistically diverse, randomly selected, resemble closely, vaguely defined, politically incorrect, point + rightly, perfectly understandable, represent + graphically, behave + differently, interestingly enough, comment + briefly, summarize + briefly, hardly surprising, widely known, evolve + constantly, closely intertwined, truly representative, overlap + partially, test + empirically, extremely rare, still perfectible, closely related*

Examples of *advmod* dependencies with 0 > MI score > 1

- *clearly negative, clearly described, important enough, measure + typically, represent + directly, very theoretical, much important, less striking, realize + even, remain + especially, rather neutral, find + usually, especially negative, even pertinent, belong + usually, quite + relevant, probably easy, express + commonly, particularly frequent, very surprising, plan + obviously, express + naturally, undoubtedly important, allow + generally, still common, slightly often, use + generally, focus + especially, obviously different, really difficult, previously seen, however significant, widely considered, often described, use + differently, highly likely, think + probably, discuss + frequently, much plausible, influence + clearly, very varied, suggest + already, previously said, provide + interestingly, often considered, previously suggested, certainly interesting, already said, happen + regularly, still confronted, very frequently, describe + simply, already identified, translate + differently, influence + partly, combine + typically, understand + immediately, focus + only, define + easily, analyze + correctly, very critical, confirm + clearly, use + mostly, rely + strongly, refer + simply, very formal, entirely true, obviously possible, first attempt, judge + easily, occur + only*

dobj dependencies

F(2,98) = 8,636, p = 0,000358, eta squared = 0,1538

	Estimate	Std. Error	t value	Pr(>|t|)
C1 – B2	0.18	0.09	1.962	0.12338
C2 – B2	0.59	0.14	4.156	< 0.001 ***
C2 – C1	0.40	0.13	3.175	0.00541 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05
(Adjusted p values reported -- single-step method)

UCL0007-LING-01 (mean MI = 1.02)	MI	UCL0020-LING-02 (mean MI = 2.99)	MI
see + appendix	5.43	pursue + career	7.83
dedicate + article	4.80	place + emphasis	7.80
cover + span	4.19	paint + picture	7.72
count + compound	3.67	project + persona	7.70
encounter + word	3.56	stigmatize + variety	7.57
compare + result	3.07	play + role	6.85
distinguish + kind	2.70	say + least	6.59
describe + process	2.16	obscure + fact	6.40
pick + term	2.09	project + image	6.12
say + word	1.85	do + justice	5.95
encompass + process	1.85	espouse + view	5.95
publish + result	1.81	assume + persona	5.92
use + approach	1.71	adopt + stance	5.81
shorten + word	1.64	construct + identity	5.48
draw + figure	1.54	conduct + study	5.44
keep + one	1.43	test + veracity	5.22
fit + scope	1.36	assemble + corpus	4.92
perceive + it	1.24	overemphasize + aspect	4.88
compare + diagram	1.16	follow + procedure	4.22
have + suffix	1.11	make + reference	4.14

Negative MI scores

- *define + source, have + change, include + increase*

- *Algeo (1991: 3-14)* **defines** *six basic etymological* **sources** *for new words: creating, borrowing, combining, shortening, blending and shifting and a seventh for new words whose source is unknown.* (UCL0007-LING-01)

Syntactic complexity

	B2		C1		C2		Between-group comparisons
	Mean	SD	Mean	SD	Mean	SD	
C/T	1.73	0.21	1.77	0.21	1.66	0.19	$F(2,98)=1.606$, p= 0.206
DC/T	0.63	0.19	0.69	0.17	0.60	0.13	$H(2,98)=1.607$, p= 0.206
DC/C	0.36	0.07	0.38	0.06	0.36	0.05	$F(2,98)=1.74$, p= 0.181
MLC	**10.67**	**1.22**	**11.16**	**1.66**	**11.50**	**1.12**	**$F(2,98)=1.436$, p=0.243**
VP/T	2.07	0.29	2.11	0.32	2.01	0.25	$H(2,98)=0.74799$, p= 0.688
CN/T	2.55	0.64	2.73	0.61	2.70	0.50	$H(2,98)=2.2303$, p= 0.3279
CN/C	**1.47**	**0.26**	**1.54**	**0.31**	**1.63**	**0.25**	**$H(2,98)=3.1148$, p=0.2107**

- No statistically significant difference

Lexical diversity

	B2		C1		C2		Between-group comparisons
	Mean	SD	Mean	SD	Mean	SD	
RTTR	**11.41**	**1.72**	**11.46**	**1.68**	**12.72**	**1.38**	**F(2,98)=2.98, p = 0.09**
LV	0.30	0.06	0.30	0.06	0.35	0.08	H(2,98)=5.29, p = 0.07
CVV1	**4.75**	**0.97**	**4.80**	**0.82**	**5.27**	**0.66**	**F(2,98)=1.98, p = 0.16**
VV2	0.08	0.01	0.08	0.02	0.09	0.02	H(2,98)=2.37, p = 0.31
NV	0.27	0.06	0.26	0.06	0.32	0.08	H(2,98)=6.21, p = 0.04
AdjV	0.07	0.01	0.07	0.01	0.09	0.02	H(2,98)=5.16, p = 0.08
AdvV	0.02	0.01	0.02	0.01	0.02	0.01	H(2,98)=4.48, p = 0.11

Alpha set at 0.05/7 = 0.007

- No statistically significant difference

Lexical sophistication

	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

Alpha set at 0.05/5 = 0.01

- No statistically significant difference

Summary

- Syntactic complexity **X**

- Lexical diversity **X**

- Lexical sophistication **X**

- Phraseological diversity **X**

- Phraseological sophistication I: academic collocations **(√)**

- Phraseological sophistication II: MI scores **√√**

CONCLUSION

Phraseological complexity

- Dimension of L2 writing quality

- Linguistic competence development from upper-intermediate to very advanced proficiency level is for the most part situated in the phraseological dimension, and not in syntactic or lexical complexity (see also Paquot & Naets, 2015)

Context-sensitive measures

- "It is (…) essential that complexity accounts for context" (Rimmer, 2009: 31)
- Register and genre
  - Operationalize the complexity of L2 language by how well it uses the phraseological units and lexico-grammatical characteristics of the norms of its reference genre (cf. Ellis et al, 2013)
- Role of the reference corpus (cf. Paquot & Naets, 2017)

Work in progress I

- **Types of word combinations**
  - Lexical bundles, P-frames, etc.
- **Other measures**
  - Phraseological diversity
    - More sophisticated measures than TTRs (cf. Jarvis & Daller, 2013)
  - Phraseological sophistication I
    - New list of academic collocations?
  - Phraseological sophistication II
    - Other statistical measures (Delta P)

47

Work in progress II

- **Replication studies**
  - L2 language across modes, tasks and genres (Paquot & Naets, 2015; Paquot & Naets, 2017b; future work with V. Brezina & D. Gablasova on the Trinity Lancaster Spoken Learner Corpus)

- **Properties**
  - Diversity, sophistication, … ?

- **Cross-linguistic validity**
  - L2 Dutch (FWO project in collaboration with A. Housen)

Implications for language assessment

- Automated techniques to investigate the phraseological competence of EFL learners (e.g. Crossley, Cai & McNamara, 2012; Bestgen & Granger 2014; Granger & Bestgen, 2014, Crossley, Salsbury & McNamara, 2014).
- Phraseological complexity should feature more prominently in language proficiency descriptors and second language assessment rubrics (Paquot, to appear 2018)
  - Idiom principle (Sinclair, 1991)
  - Phraseology: a challenge to language learners
  - Differentiate /b/ the most advanced proficiency levels
- Augment the set of linguistic indices used to automatically score L2 productions

49

Phraseological complexity and the Common European Framework of References for Languages (CEFR)

- The CEFR needs updating to account for recently accumulated knowledge on how lexis and grammar are intertwined.
  - Section 5.2.1 on linguistic competence
  - Not a single mention of phraseology, collocations, formulaic sequences in the *Structured Overview of all CEFR scales* (Council of Europe, 2001)
- A better understanding of the development of phraseology and lexico-grammar in learner language could balance out the focus on education or cognitive development that has so far served to identify C1 and C2 levels (cf. Alderson, 2007; Hulstijn, 2015).

THANK YOU!

- Paquot, M. (2017). The phraseological dimension in interlanguage complexity research. Second Language Research. Second Language Research. 10.1177/0267658317694221

- Paquot, M. (to appear 2018). Phraseological competence: a useful toolbox to delimitate CEFR levels in higher education? 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)

- Paquot & Naets (2017) The role of the reference corpus in studies of EFL learners' use of statistical collocations. Paper presented at ICAME, Prague, 25-28 May 2017.

Check out!

- **The Learner Corpus Association**
  - www.learnercorpusassociation.org

- **The International Journal of Learner Corpus Research**
  - General editors: Marcus Callies & Magali Paquot
  - John Benjamins Publishing