
ALWAYS
LEARNING

The use of academic collocations in essays in a test of academic English

Dr Veronica Benigno (Pearson)
Jeremy Hancock (Pearson)
Katarzyna Pawlak (Pearson)
Dr Olivier Kraif (Université de Grenoble)

LTRC 2014
4 - 6 June 2014,
VU University Amsterdam, The Netherlands

Outline

SECTION I: Introduction

SECTION II: Data

SECTION III: Method and tools

SECTION IV: Results

SECTION I: Introduction

Problem statement

- Research has shown vocabulary knowledge is a good predictor of proficiency and that a relationship exists between vocabulary measures and communicative skills, reading in particular

- However, studies have mainly focused on vocabulary size and use of single words rather than on phraseological units

Research goal

What is the relationship between **USE of general and academic collocations** in 50,000 essays produced by test takers of PTE Academic (Pearson Test of English Academic) and **PROFICIENCY level?**

Research questions (1, 2, 3, and 4)

RQ1: What is the relationship between USE of general and academic collocations and PROFICIENCY?

USE is defined by:

1) Number
2) Variety
3) Syntactic patterns
4) L1 frequency

PROFICIENCY is defined by:

Test takers' CEFR level as measured by PTE Academic

Collocations

A type of phraseological unit (Wray, 2002)

Ready-made combinations of words that are expected to come together

heavy smoker (EN);

**persistent smoker (IT fumatore accanito); *strong smoker (GE starker Raucher)*

Difficult to master even for advanced learners (Laufer & Waldman, 2011)

Part of communicative competence (Henriksen, 1999)

Indicators of vocabulary depth (Read, 2004; Schmitt, 2010)

Discriminating language use by NS and NNS (Benigno & Vedder, 2013; Laufer & Waldman, 2011;)

SECTION II: Data

Data - overview

Corpus of essays

50,000 essays produced by test takers of PTE Academic – A2 to C2

Reference lists

*a) Academic Collocations List –ACL- (*Ackermann & Chen, 2012)

- Includes 2,469 academic collocations extracted from PICAE

b) General Collocations List –GCL-

- Includes 110,000 general collocations extracted from Longman Dictionary of Contemporary English (LDOCE) and Longman Collocations Dictionary and Thesaurus (LCDAT)

L1 corpora

a) *Longman Corpus Network (LCN)*
- Consists of 330 million words (tokens)
- Includes British and American spoken and written texts
- Balanced and representative of general English

b) Pearson International Corpus of Academic English (PICAE)
-Consists of 37 million words (tokens)
-Includes academic written (13%) and spoken (87%) texts
-Sourced from the web (52%) + other written sources

Academic collocations (examples)	Conduct research, address issue, gain insight, significant contribution
General collocations (examples)	Make decision, ask question, health care, take time

SECTION III: Method and tools

Step 1: Corpus compilation and data treatment

Corpus of essays

50,000 essays from CEFR levels A2 to C2

Data pre-treatment

- Automatic spell check to correct test takers' typos and errors

  > Use of Microsoft Word Spell Checker (piloted on a subset)

- Lemmatization and POS-tagging using Tree tagger
- Collocations occurring in the prompt removed from the counting to exclude any collocations that might be lifted

Step 2: Extraction of collocations and matching

Extraction

Selected parameters

> span 4 > order sensitive > no proper nouns > POS restriction (Nouns, Verbs, Adjectives and Adverbs)

Matching

Collocations extracted from the **corpus** were matched against the two **reference lists***

** Collocations which were in both reference lists were removed from the GCL to keep the two lists distinct.*

The corpus of essays

Corpus of essays

50,000 essays from CEFR levels A2 to C2
total: more than 11,000,000 words

PTE Academic :

scores are reported on the **Global Scale of English (GSE)**
GSE score range: 10 - 90

CEFR level	Equivalent GSE score range
A2	30 - 42
B1	43 - 58
B2	59 - 75
C1	76 - 84
C2	85 - 90

Selection criteria for the essay

- Minimum overall score CEFR A2 (GSE 30)
- Minimum writing score CEFR A1 (GSE 22)
- Essays 170 to 330 words
- Minimum time spent on essay 5 mins (max is 20 mins)
- Essay score greater than 0
- NS removed (birth country, citizenship, language at home)

‘Write essay’ item in PTE Academic

Task: persuasive or argumentative essay of 200-300 words on a given topic (time allowed: 20 min.; scoring: partial credit)

Traits:

- Content
- Formal Requirement
- Development, Structure and Coherence
- Grammar
- General Linguistic Range
- Vocabulary Range
- Spelling

Sample PTE Academic essay prompt

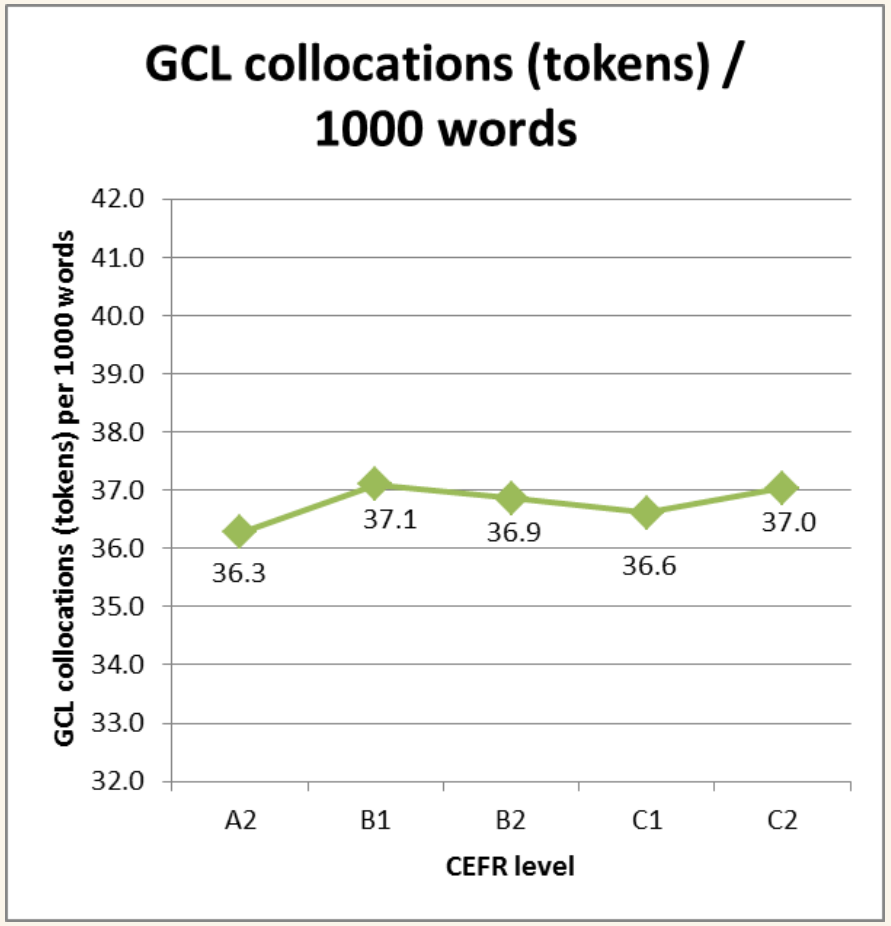
You will have 20 minutes to plan, write and revise an essay about the topic below. Your response will be judged on how well you develop a position, organize your ideas, present supporting details, and control the elements of standard written English. You should write 200-300 words.

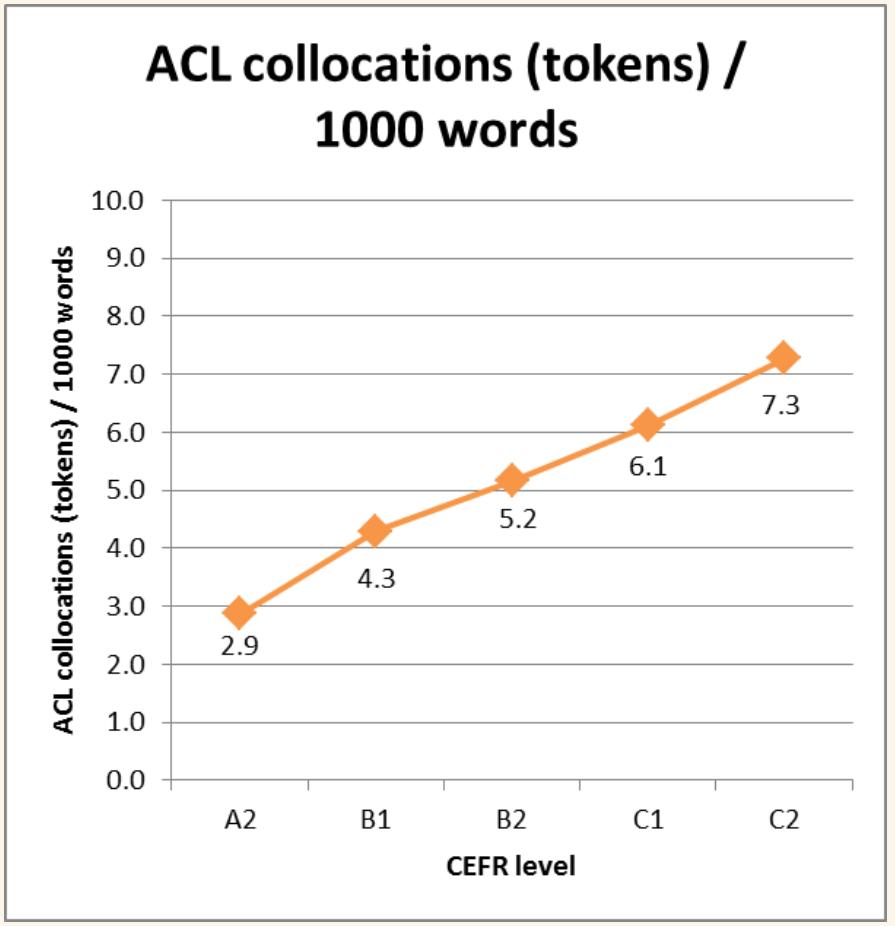
Tobacco, mainly in the form of cigarettes, is one of the most widely-used drugs in the world. Over a billion adults legally smoke tobacco every day. The long term health costs are high - for smokers themselves, and for the wider community in terms of health care costs and lost productivity.

Do governments have a legitimate role to legislate to protect citizens from the harmful effects of their own decisions to smoke, or are such decisions up to the individual?

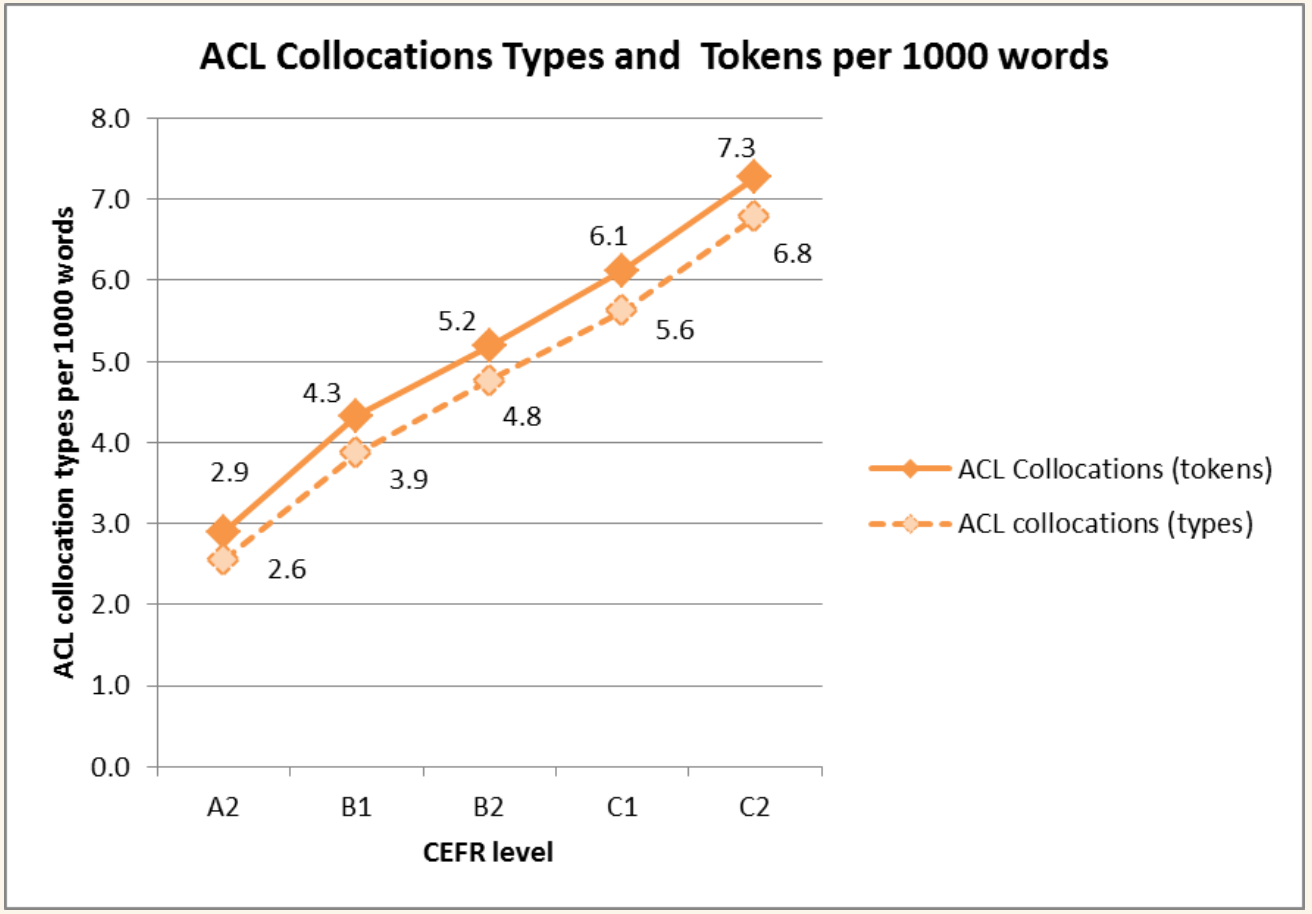
SECTION IV: Results

RQ1 – Relationship between number of collocations and proficiency – GCL and ACL

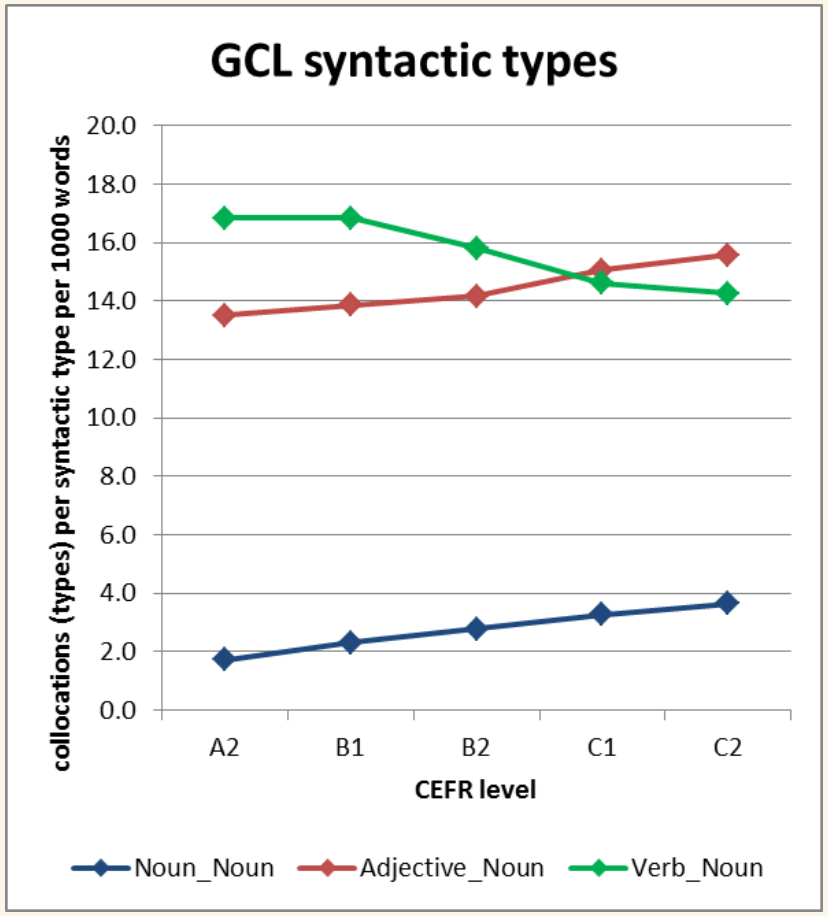


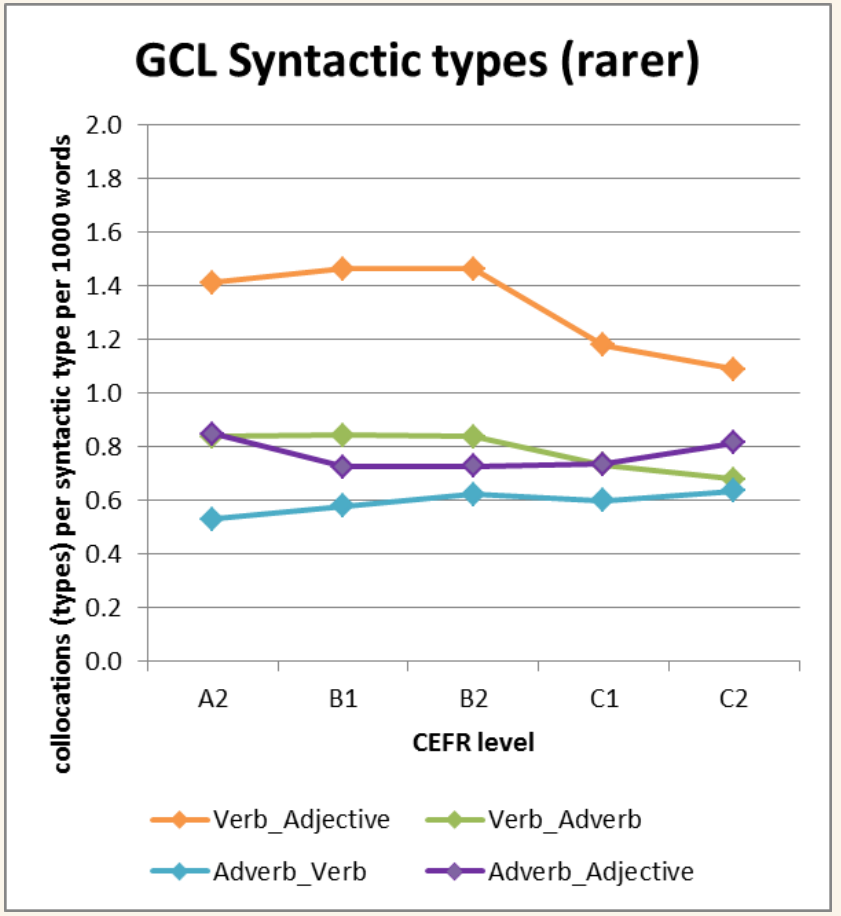


RQ2 – Relationship between variety of collocations and proficiency – ACL



RQ3 – Relationship between syntactic patterns and proficiency – GCL tokens

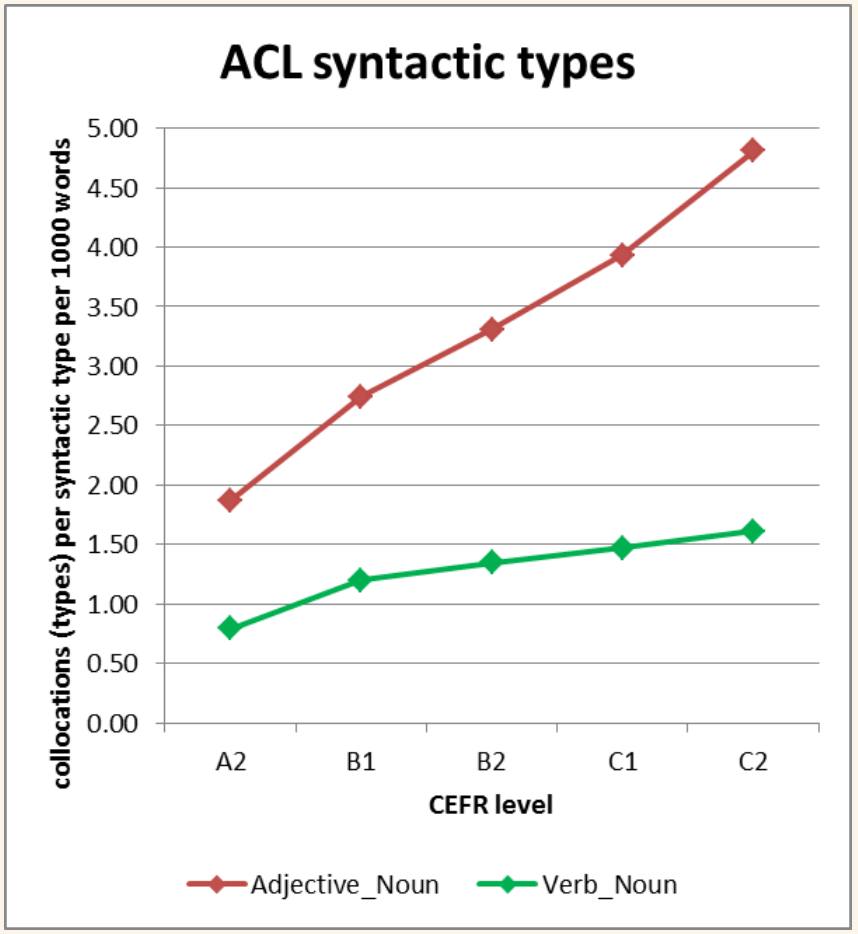


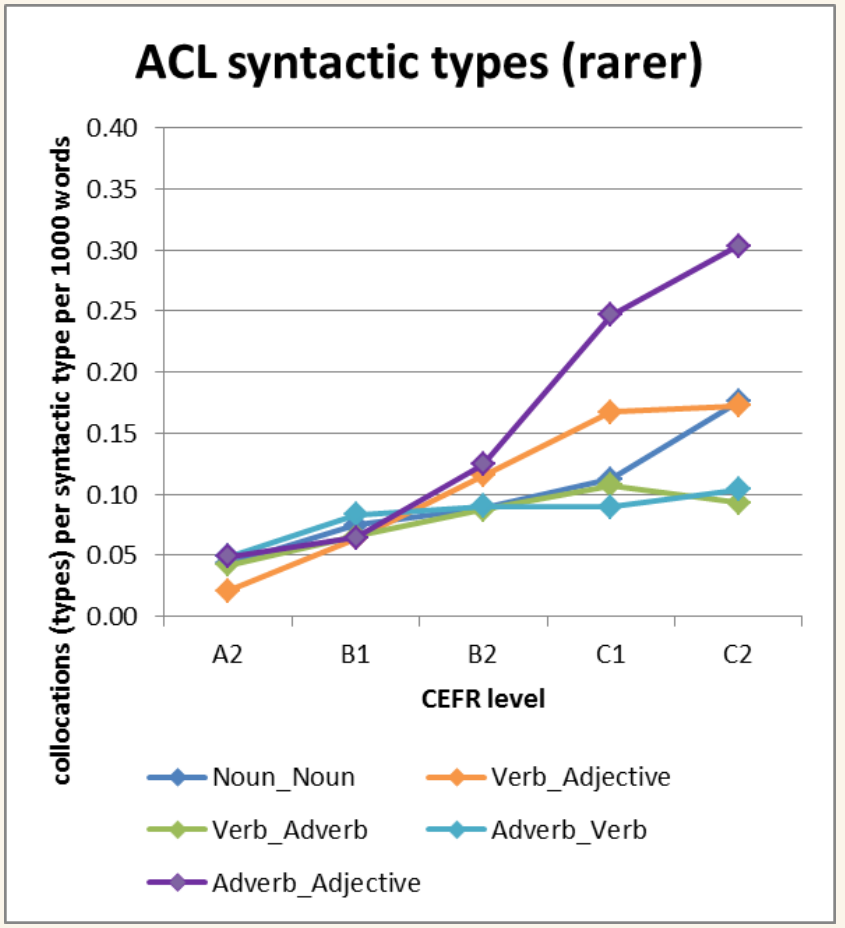


RQ3 – Relationship between syntactic types and proficiency – GCL examples

Noun Noun:	*health problem, world war, safety measure*
Adjective Noun:	*serious damage, dark side, personal opinion*
Verb Noun:	*make decision, solve issue, pay attention*
Verb Adjective:	*go wrong, find difficult, prevent damage*
Adverb Adjective:	*absolutely necessary, easily available, highly rewarding*
Adverb Verb:	*completely agree, strongly believe, badly damage*
Verb Adverb:	*push hard, work hard, cut down*

RQ3 – Relationship between syntactic patterns and proficiency – ACL tokens

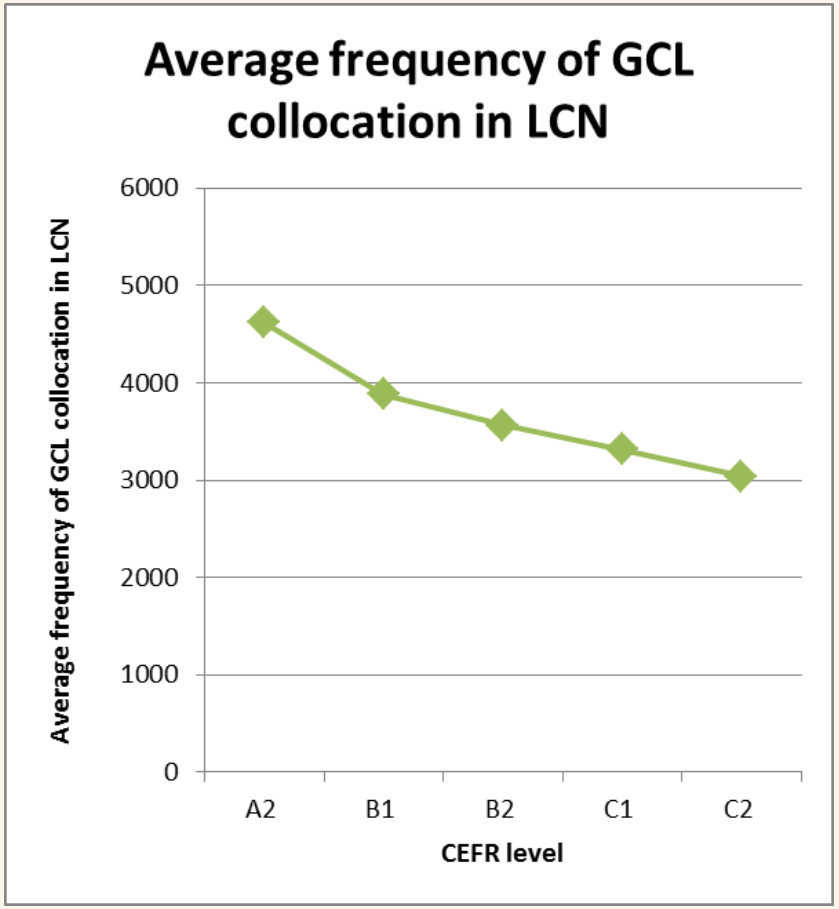


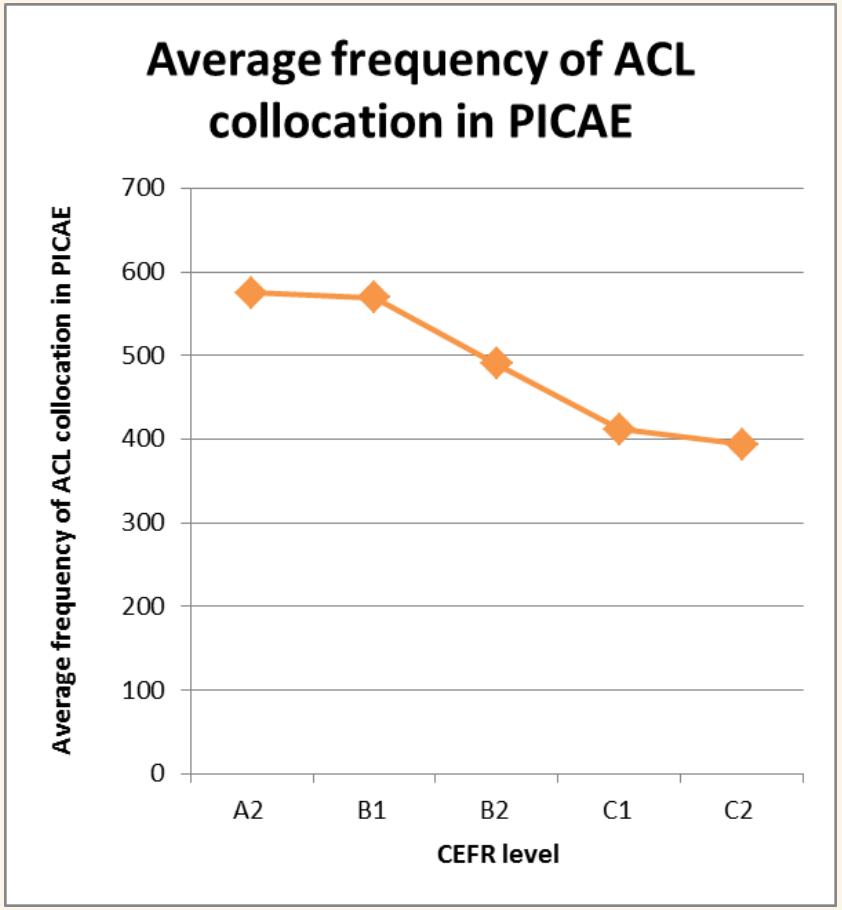


RQ3 – Relationship between syntactic types and proficiency – ACL examples

Noun Noun:	*business transaction, government policy, research methodology*
Adjective Noun:	*significant contribution, economic growth, integral part*
Verb Noun:	*gain insight, face challenge, pose threat*
Verb Adjective:	*become evident, prove useful, remain unchanged*
Adverb Adjective:	*increasingly important, widely available, rapidly growing*
Adverb Verb:	*significantly increase, directly affect, briefly discuss*
Verb Adverb:	*rely heavily, grow rapidly, vary widely*

RQ4 – Relationship between corpus frequency of collocations and proficiency – GCL and ACL





Conclusions

Use of collocations is positively related with proficiency

- **Number** of collocations used by test takers increases with their proficiency level, but only for academic collocations
- **Variety** of collocations used increases with proficiency for academic collocations at a similar rate as tokens
- For **general** collocations, two **syntactic types** (NN, AN) are positively related to proficiency
- For **academic** collocations, most **syntactic types** are positively related to proficiency (with NN, AN being stronger indicators)

ADVA type (e.g. *increasingly important, widely available etc.)* shows the potential for discriminating between proficiency levels for academic English

- Less proficient learners tend to use high-**frequency** collocations whereas advanced learners use more rare combinations

Contribution

Insight into how analysis of collocations could help better define and assess vocabulary

- Use of collocations (as a depth measure) in addition to standard measures focusing on single words only

Future research direction

Investigate to what extent test takers' demographic information, e.g. L1, affects use of collocations

- Does test takers' L1 play a role in use of collocations at different proficiency levels?

References

- **Ackermann, K., & Chen, Y.** (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. Journal of English for Academic purposes, Vol. 12, Issue 4.
- **Benigno, V., & Vedder, I.** (2013). La ricorrenza del lessico di base in produzioni scritte di italiano L2 e L1. Linguistica e Filologia, vol. 33, 59-84.
- **Henriksen, B.** (1999). Three dimensions of vocabulary development. Studies in Second Language Acquisition, 21(2), 303–317.
- **Laufer, B., & Waldman, T.** (2011). Verb-noun collocations in second-language writing: A corpus analysis of learners' English. Language Learning, 61(2), 647-672.
- **Lésniewska, J., & Witalisz, E.** (2007). Cross-linguistic influence and acceptability judgments of L2 and L1 collocations: A study of advanced Polish learners of English. EUROSLA Yearbook 7, 27-48.
- **Martinez, R. & Schmitt, N.** (2012). A phrasal expression list. Applied Linguistics, 33(3), 299-320.
- **Read, J.** (2000). Assessing Vocabulary. Cambridge: Cambridge University Press.
- **Schmitt, N.** (2010). Researching vocabulary – A vocabulary research manual. Basingstoke: Palgrave Macmillan.
- **Schmitt, N., Jiang, X., & Grabe, W.** (2011). The relationship between the amount of vocabulary known in a text and reading comprehension. The Modern Language Journal, 95(1), 26-43.
- **Wray, A.** (2002). Formulaic language and the lexicon. Cambridge: Cambridge University Press.

Thank you!