An Academic Collocation List

Kirsten Ackermann
BALEAP 2011, Portsmouth

ALWAYS LEARNING PEARSON

Contents

Introduction

1

Motivation

- Significance: collocations are essential for successful language processing and language use
- Dilemma: collocations are instantly recognized by native speakers, but remain difficult for learners to acquire and use properly
- Proficiency: collocations make fluent language possible, support comprehension, create a register

Objective

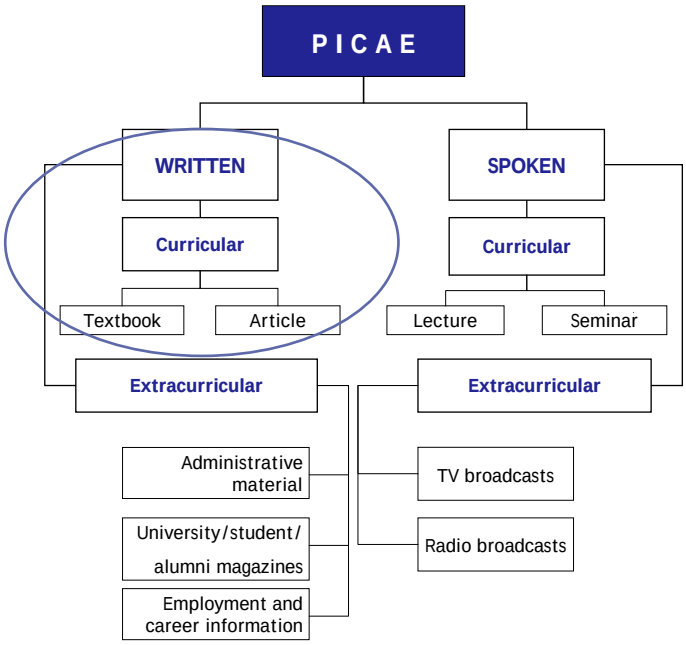
- To compile a list of the most frequent and pedagogically relevant collocations in written academic English in order to devise efficient learning, teaching and assessment resources

Tool

- Pearson International Corpus of Academic English (PICAE)

Pearson International Corpus of Academic English (PICAE)

2



PICAE: Written curricular component

- 333 documents
  - From 4 academic disciplines: *Humanities, Social Science, Natural and Formal Science, Professions and Applied Science*
  - Covering 28 academic subjects: *7 per academic discipline*

PICAE: Academic disciplines and subjects

Humanities		Social Science		Natural / Formal Science		Professions / Applied Science	
Subject	Tokens	Subject	Tokens	Subject	Tokens	Subject	Tokens
History	946,707	Anthropology	413,237	Earth Sciences	1,343,723	Architecture	167,074
Linguistics	855,128	Archaeology	184,089	Chemistry	1,502,277	Business	1,644,180
Literature	1,562,046	Cultural studies	861,656	Physics	662,054	Education	405,202
Art incl. Music	728,532	Gender studies	520,395	Computer sciences	1,124,097	Engineering	1,134,950
General academia	627,951	Politics	1,090,800	Mathematics	295,565	Health sciences	1,429,679
Philosophy	602,233	Psychology	1,560,745	Biology	858,597	Media studies	1,500,485
Religion	198,165	Sociology	1,832,588	Ecology	239,787	Law	1,962,002
Total	5,520,762	**Total**	6,463,510	**Total**	6,026,100	**Total**	8,243,572

Methodology

3

Overview

Step One – Douglas Biber's quantitative analysis of PICAE

Result: list of c.130,000 collocations and extended phrases

Step Two – Manual vetting of initial collocation list

Result: list of c.3,500 collocations

Step Three – Expert judgment

Result: list of c.3,000 pedagogically relevant collocations

Stage One: Quantitative analysis of the corpus data
Target output

Initial academic collocation list

With 'collocation' defined as:

- A single word that tends to co-occur in the span of ±3 words from the target word
- ≥1 per million words
- In ≥5 different texts
- Mutual Information score ≥3
- T-score ≥2

Stage One: Quantitative analysis of the corpus data
Computational method

A) Complete reference list

- Words appearing more than 12 times in the corpus

B) Reference list of content words

- Content words (nouns, verbs, adjectives, adverbs)
- Frequency of ≥5 per million words
- Distribution in ≥5 different texts
- Excluding General Service List headwords
- Using 'stop list' to exclude frequent function words with purely grammatical meaning

Stage One: Quantitative analysis of the corpus data
Computational method

C) Initial list of c.130,000 academic collocations

Information for each collocate includes:

- Location relative to the target word (-3, -2, -1, +1, +2, +3)
- Normed overall frequency (per million words)
- Normed frequency (per million words) in each academic discipline
- Number of texts it occurs in
- Mutual Information score
- T-score

Pre-collocate	Academic Word	Post-collocate	Most Freq Position	Normed Freq per million	MI score	t-score	# of texts	PAS	HM	SS	NFS
services	provided		-1	8.12	6.39	13.29	45	16.56	0.21	8.67	3.22
	rationale	for	1	8.12	5.93	13.23	79	11.14	7.34	8.31	4.42
	cultural	xxx political	2	8.08	4.82	12.94	55	5.00	5.66	16.44	5.43
	gender	identity	1	8.08	7.17	13.32	32	1.71	7.76	18.79	5.43
other xxx	including		-2	8.08	3.72	12.40	102	9.56	5.45	8.67	7.84
provides	information		-1	8.08	5.52	13.12	67	6.85	3.35	9.39	12.87
to	interact		-1	8.08	3.21	11.97	69	8.57	5.45	10.48	7.24
whole	range		-1	8.08	6.52	13.27	64	8.85	6.29	12.46	3.82
european	community		-1	8.03	6.00	13.17	42	17.56	1.89	5.42	3.42
	experimental	data	1	8.03	6.87	13.27	23	3.43	1.47	0.72	28.96
social	issues		-1	8.03	3.99	12.54	59	6.28	2.31	19.51	3.22
can xxx	measured		-2	8.03	4.96	12.95	71	5.42	4.40	4.70	18.91
	remote	sensing	1	8.03	14.45	13.38	10	0.14	0.00	0.00	35.80
may	affect		-1	7.99	5.45	13.04	71	12.28	4.19	5.96	7.84
	information	society	1	7.99	4.61	12.79	12	21.70	0.00	3.79	1.01

Step Two: Manual vetting of initial collocation list

1. Preparation

- Tagging initial collocation list with OpenNLP using a simplified set of tags
- Converting tagged output into specified format

Pre-collocate	Pre-C tag	Academic word	AW tag	Post-collocate	Post-C tag
key	adj	concepts	n		
from xxx	prep xxx	region	n		
different	adj	areas	n		
		culture	n	xxx society	xxx n
		community	n	based	vpp
social	adj	organization	n		
future	adj	research	n		
		slightly	adv	more	adv
at xxx	prep xxx	disposal	n		
order xxx	n xxx	obtain	v		
cultural	adj	diversity	n		
		independent	adj	variables	n

Step Two: Manual vetting of initial collocation list

2. Cleaning based on quantitative parameters

- Normed frequency ≥1
- MI score ≥3
- T-score ≥4
- Dispersion ≥.2

3. POS-analysis of the cleaned list

- Filtering by POS tags to include only target combinations
  - N+V N+Vpp
  - Adj+N Adv+Adj
  - Adv+V
- Re-tagging if applicable

Step Two: Manual vetting

4. Qualitative analysis

- Judging each collocation whether to
  - include
  - discuss
  - exclude

  in the academic collocation list.

Step Three: Expert judgment

1. Each collocation is to be judged by an expert committee on its pedagogical value.
2. A decision is to be made independently by each expert whether or not to include the collocation in the final list.
3. The list will be finalised in panel discussion.

Result: An Academic Collocation List

The Academic Collocation List is a list of the most frequent and pedagogically relevant collocations in academic written English that:

- Can help students from all academic disciplines to increase their collocational competence and thus their language proficiency
- Can assist EAP teachers in their lesson planning
- Will inform test development, i.e. item writing, item type, item analysis
- Will provide a research tool for investigating the development of academic language proficiency

The Academic Collocation List
A sample

4

PEARSON

Sample I:
Frequency

Pre-collocate	Pre-C tag	Academic word	AW tag	Post-collocate	Post-C tag	Status
social	adj	policy	n			v
popular	adj	culture	n			f
local	adj	authority	n			v
wide	adj	range	n			v
		cultural	adj	studies	n	s
		previous	adj	chapter	n	d
local	adj	authorities	n			v
free	adj	movement	n			v
social	adj	security	n			s
private	adj	sector	n			v
public	adj	policy	n			v
		indigenous	adj	peoples	n	s
political	adj	economy	n			s
		widely	adv	used	vpp	v
		ethnic	adj	groups	n	s
general	adj	assembly	n			v
local	adj	government	n			s
		academic	adj	writing	n	s
		productivity	nn	growth	n	v
academic	adj	writing	n			s
public	adj	sphere	n			s
human	adj	genome	n			f/s
low	adj	income	n			c
		individual	adj	differences	n	c
resources	n	available	adj			c

PEARSON

Sample II: Collocation family 'development'

	Pre-collocate	Academic word		Academic word	Post-collocate		Pre-collocate	Academic Word	Post-collocate
v	*subsequent*	development	v	developmental	processes	?	affect xxx	development	
v	further		v		stage	?	contribute xxx		
v	*professional*		v	develop	strategies	?	ensure xxx		
s	emotional		v	*physical*	development	?	facilitate xxx		
s	*physical*		v	*professional*		?	promote xxx		
s	human		v	*subsequent*		?		develop xxx	theory
s	economic		v	*technological*		?		developed xxx	theory
s	social		v	subsequent	developments	?		developing xxx	strategies
s	historical		v	technological		?		developing xxx	theory
s	agricultural		v	fully	developed				
s	spiritual		v	highly					
s	industrial		v	originally					
s	urban								
s	*technological*								
s	regional								
c	significant								
c	normal								
c	future								

Sample III: Collocation family 'research'

	Pre-collocate	AW		AW	Post-collocate		Pre-collocate	AW
adj	considerable	research	n	research	efforts	adj	traditional	research
adj	initial		n		effort	adj	comparative	
adj	earlier		n		purposes	adj	educational	
adj	past		n		methodology	adj	experimental	
adj	original		n		evidence	adj	economic	
adj	primary		vpp/adj		published	adj	historical	
adj	extensive		vpp/adj		undertaken	adj	national	
adj	little		vpp/adj		conducted	adj	medical	
adj	major					adj	academic	
adj	basic					adj	quantitative	
adj	current					adj	qualitative	
adj	empirical							
adj	previous							
adj	future							
adj	scientific							
adj	further							
adj	recent							
p/adj	existing							
p/adj	published							
n	field							
v	undertake							
v	conducting							

Discussion

5

PEARSON

Discussion

- What about the collocations under discussion?
- Is there a preferred way of presenting the collocations?
- Are there any additional usages for the list?
- What to do with extended phrases?

1. Some included, others rejected
2. If POS is not a target combination
3. Not a linguistically complete unit
4. Degree of fixedness
5. Transparent but useful discourse organisers/referential expressions
6. Common adjectives as headwords
7. Subject specific
8. Implication/connotation

PEARSON

Thank you

kirsten.ackermann@pearson.com