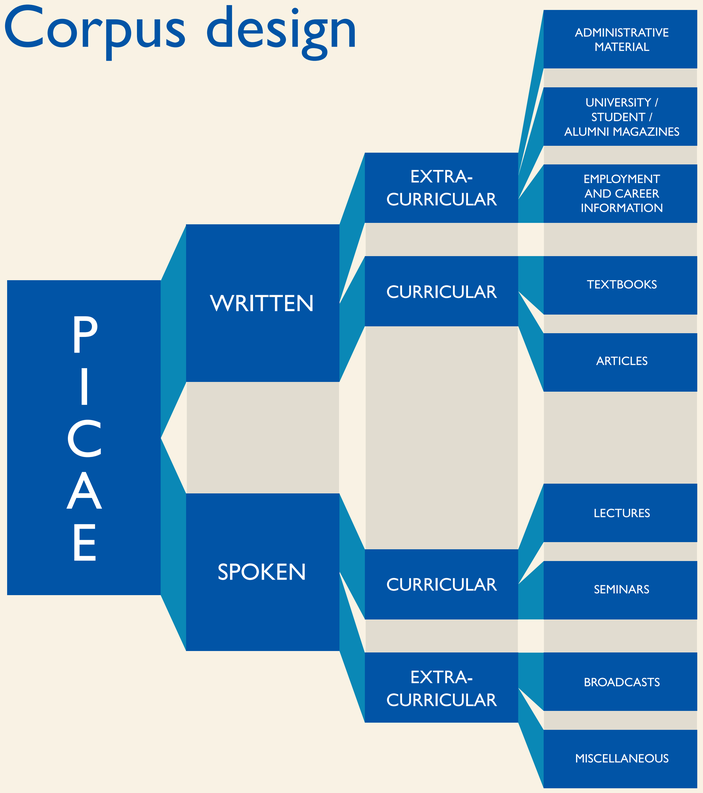
The Pearson International Corpus of Academic English (PICAE)

Kirsten Ackermann[†], John H.A.L. de Jong[†], Adam Kilgarriff[‡], David Tugwell[‡]
[†]Pearson, [‡]Lexical Computing Ltd

Introduction

PICAE comprises over 37 million words, including 13% spoken and 87% written material. The corpus is designed to reflect language that the non-native speaker will encounter in academic settings where English is the main language used. PICAE, therefore, includes both the English needed for academic work, i.e. curricular English (72%), and the English needed for various aspects of extracurricular life, i.e. extracurricular English (28%). PICAE covers five varieties of English: American, Australian, British, Canadian and New Zealand English. The corpus is used to further explore the register of academic English in order to support language teaching and assessment.

Corpus design



Component	Words
WRITTEN	32,475,526
Written Curricular	25,614,737
Textbooks	*19,627,558*
Articles	*5,987,179*
Written Extracurricular	6,860,789
Administrative	*1,165,539*
Magazines	*5,288,573*
Employment	*406,677*

Corpus components

Component	Words
SPOKEN	4,640,675
Spoken Curricular	1,027,598
Lectures	*751,203*
Seminars	*276,395*
Spoken Extracurricular	3,613,077
Broadcasts	*3,320,042*
Miscellaneous	*293,035*
PICAE Total	37,116,201

Academic disciplines and subjects in the written curricular component

Humanities		Social Sciences		Natural / Formal Sciences		Professions and Applied Sciences	
Discipline	Words	Discipline	Words	Discipline	Words	Discipline	Words
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
Arts	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

Putting PICAE to use: The Academic Collocation List

The Academic Collocation List is a list of the most frequent and pedagogically relevant collocations in written academic English discourse. The list facilitates, for example, EAP material development, item development, and validity research into the *Pearson Test of English Academic*.

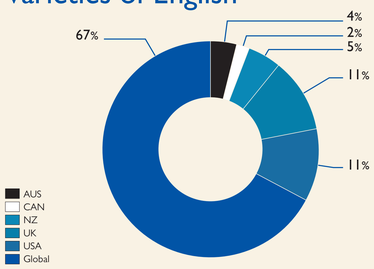
This list was derived from the written curricular component of PICAE, which comprises over 26 million words from 333 documents covering 28 major subjects from four academic disciplines: *humanities, social sciences, natural and formal sciences, professions and applied sciences*.

Academic Collocation List: Sample entries

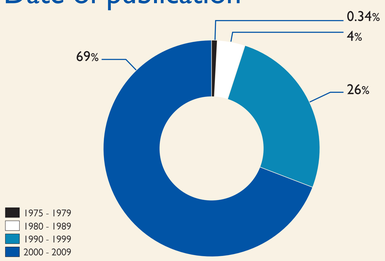
Pre-Collocate	Academic Word	Post-Collocate	Normed frequency in PICAE: per million	Applied Sciences	Humanities	Social Sciences	Natural/ Formal Sciences	MI score	t-score
	academic	discourse	1.30	0.71	2.52	1.99	0.20	5.65	5.28
	acquire	knowledge	2.56	1.57	4.82	3.43	0.80	7.66	7.51
detailed	analysis		6.78	9.71	4.61	4.88	6.84	7.19	12.20
readily	available		5.88	7.28	3.98	3.79	8.04	8.39	11.41
	cognitive	skills	2.06	1.86	0.63	5.06	0.40	6.64	6.71
	competitive	market	3.50	7.85	1.26	1.99	1.21	7.98	8.80
increasingly	complex		2.51	2.14	1.68	3.79	2.41	6.12	7.38
informed	consent		7.36	6.00	4.61	5.96	13.48	11.45	12.80
vary	considerably		2.87	3.00	2.52	2.35	3.62	10.49	7.99
	disclose	information	1.08	2.28	0.63	0.36	0.60	7.92	4.88
	distinct	ways	1.12	0.71	1.68	1.45	0.80	5.26	4.87
cultural	diversity		5.65	4.43	1.05	14.27	2.21	7.08	11.14
	dominant	position	6.28	15.85	1.47	3.61	0.40	7.31	11.76
	empirical	evidence	4.62	5.42	3.35	8.13	0.80	7.62	10.10
	environmental	change	13.24	1.00	0.21	0.90	56.72	7.28	17.07
became	evident		1.39	2.14	0.42	1.81	0.80	6.96	5.52
distinctive	features		3.19	2.28	3.56	5.78	1.21	8.69	8.41
	homogeneous	group	1.03	0.86	0.63	1.81	0.80	6.32	4.74
	individual	characteristics	1.93	1.86	1.05	3.97	0.60	5.01	6.35
	integral	part	9.51	12.28	9.43	10.66	4.42	8.18	14.51
further	investigation		3.77	4.14	5.24	2.89	2.82	7.02	9.09
vast	majority		11.31	13.13	7.96	15.53	7.24	11.03	15.87
	negative	effects	3.59	7.00	0.21	3.79	1.81	6.40	8.84
naturally	occurring		8.89	2.14	4.19	5.06	27.15	12.02	14.07
social	policy		57.53	13.42	2.52	211.71	0.80	6.60	35.43
	primary	sources	5.25	7.00	12.79	0.36	1.01	7.53	10.76
fundamental	principles		4.44	7.57	3.35	4.15	1.41	7.59	9.90
	radically	different	4.67	1.71	9.01	6.14	3.02	7.76	10.15
	randomly	chosen	1.71	1.43	0.42	0.54	4.63	10.82	6.16
wider	range		5.79	6.85	5.45	5.78	4.63	7.68	11.30
key	role		11.98	10.85	1.47	27.82	6.03	6.45	16.15
	salient	features	1.17	1.43	1.26	1.08	0.80	9.03	5.09
private	sector		23.20	44.97	2.52	28.90	6.03	9.95	22.71
	significant	impact	4.67	8.71	1.89	4.70	1.61	6.24	10.06
increasingly	sophisticated		1.53	2.57	1.05	0.90	1.21	8.43	5.81
public	sphere		14.99	33.83	4.61	12.46	1.21	8.88	18.24
	strategic	management	7.63	23.27	0.42	0.54	0.40	8.44	13.00
	subsequent	chapters	3.14	6.71	1.26	2.35	0.80	8.69	8.35
	technological	advances	2.24	4.00	1.05	2.17	1.01	10.89	7.07
	vital	part	2.29	2.00	2.52	3.43	1.21	5.97	7.03

Lexical Computing

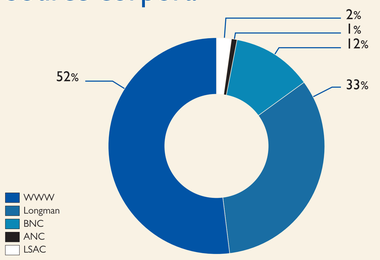
Varieties of English



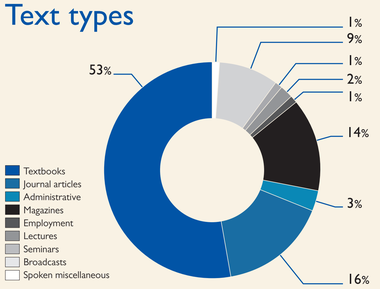
Date of publication



Source corpora



Text types



References
Kilgarriff A. & Grefenstette, G. (2003). Introduction to the Special Issue on Web as Corpus. Computational Linguistics, 29 (3): 333-348.
Kilgarriff A., Rychly, P., Smrz, P. & Tugwell, D. (2004).The Sketch Engine. Proceedings of Euralex. Lorient, France, July: 105-116.

More information
Email: pltsupport@pearson.com
Web: www.pearsonpte.com