The Pearson International Corpus of Academic English (PICAE)

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

Abstract

This paper introduces the Pearson International Corpus of Academic English (PICAE) compiled at Pearson in collaboration with Lexical Computing Ltd. As part of the development programme for Pearson Test of English Academic (PTE Academic), it was decided to compile a reference corpus of academic English to inform test development and further investigate the development of academic language proficiency in non-native English speakers. PICAE aims to reflect the register of academic English by including curricular English as found in lectures, seminars, textbooks and journal papers as well as extracurricular English that students encounter on campus from university administration to transcripts of broadcasts. The corpus comprises spoken and written data from five major English-speaking countries. PICAE has been cleaned, lemmatised and POS-tagged and is available for research.

1. Introduction

As part of the development programme for Pearson Test of English Academic (PTE Academic), it was decided in 2007 to compile an academic corpus that would comprise spoken and written data from five major English-speaking countries in order to support the objective to ground PTE Academic on an accurate representation of the English that students will need to understand and produce in academic settings where English is the language of instruction. As a reference corpus PICAE informs test development and provides a research tool for investigating the development of academic language proficiency.

This paper introduces the Pearson International Corpus of Academic English (PICAE), a corpus of over 37 million words. It contains written and spoken English and covers five major varieties of the language. The corpus includes curricular English as found in lectures, seminars, textbooks and journal papers. It also samples extracurricular English that students will encounter, from university administration to transcripts of radio broadcasts. It is mainly sourced from the web. This paper describes the three phases of the creation of PICAE: design, collection and encoding. It also provides an overview of the current composition of the corpus in charts and tables and briefly introduces one research project based on the corpus.

2. Design

Academic English is a formal register of English used in academic settings for instructional and procedural purposes. The register includes the language features that are represented across disciplines as well as the language features specific to individual academic disciplines. Knowledge of academic English is needed by all students for long-term academic success. Gillett & Wray (2006) show that teaching academic English empowers students to succeed in their academic careers. Bailey (2009) adds that academic English is not only required for successful academic achievement, but is “likely to be as necessary for success outside and/or beyond school in the

professional and business worlds students will encounter". When defining the academic register, it should, thus, not be limited to daily communication in academic settings, but also involve the competence to understand and produce any written academic material, to participate actively in lectures, seminars and tutorials, and to organise academic life on campus. Consequently a broader definition of academic English was adopted in the compilation of PICAE. The corpus comprises the English used in academic productions such as journal articles and lectures as well as the English students need in order to participate fully in academic life on campus, e.g., student journals, career information. We refer to the former as curricular English and to the latter as extracurricular English in this article. Together they form the academic English represented in PICAE.

The curricular material (72%) includes a wide range of academic subjects covering the four main academic disciplines, namely humanities, social science, natural and formal science, and professions and applied sciences. It also comprises lectures, seminars, textbooks and journal articles at undergraduate as well as postgraduate levels. The extracurricular material (28%) includes university administrative material, university/student/alumni magazines, employment and career information as well as TV and radio broadcasts. Figure 1 provides an overview of the hierarchical design of PICAE.

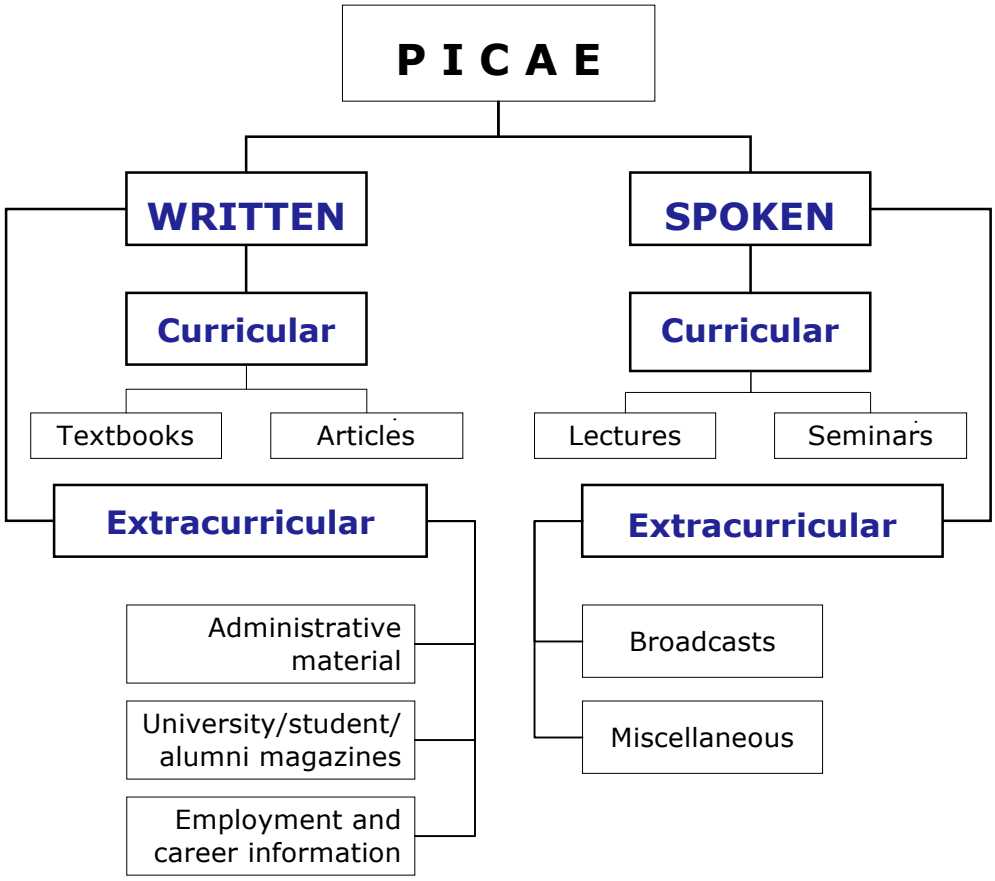

Figure 1: Design of PICAE

Furthermore, PICAE includes American, Australian, British, Canadian and New Zealand English to reflect the varieties of English non-native students are most likely to encounter when studying at university where English is the main language.

3. Collection

PICAE data was gathered from five different sources to give a total of over 37 million words:

- 19.6 million words from the World Wide Web
- 12.1 million words from Longman Higher Education textbooks
- 0.7 million words from the Longman Spoken American Corpus
- 4.4 million words from the British National Corpus
- 0.4 million words from the American National Corpus

3.1 World Wide Web

The principal source was the World Wide Web and material was collected from it in two ways according to the text type collected: (1) manually identifying suitable documents, and downloading them; (2) identifying suitable websites and downloading the whole site. An additional method was initially considered, which was to use a set of search terms associated with academia as input to a search engine such as Google or Yahoo and automatically gather all the relevant pages. Investigation of this method, however, led us to believe that the amount of work involved in ensuring quality and relevance of the data, as well as accurately categorizing the documents found, was too great to make it worthwhile.

For textbooks, the most common source was PDF versions of published textbooks made available online by their publishers. Such collections of documents are generally presented in a standard format, with a standard representation of author, title and date of publication, making it possible to semi-automate the process of collection. Similarly, for journal articles the most common source was PDF versions collected from freely accessible online journals. One manual task was the assignment of the correct academic discipline to the document. Documents in PDF format were converted into text files using the unix utility 'pdftotext'. In line with the practice of the BNC, we limited the amount of text taken from any one written academic document to 40,000 words in order to increase the spread of vocabulary and constructions in this section of the corpus.

For extracurricular written texts associated with university life and administration, our main sources were university and college websites. These sites had a key advantage for our purposes that they could be unambiguously associated with a region, thus helping to ensure that we maintained a representative spread in the varieties of English we were interested in. Administrative material was either collected in HTML form or whole PDF documents were downloaded and converted. Texts representative of general student life were found in the form of student magazines, alumni magazines and university journals that could be downloaded from online archives. Once again the advantage of downloading multiple documents from a single site was that the process could be semi-automated, with contents and dates being assigned at the same time, and a significant volume of material collected with a comparatively low level of human investigation and input.

Not surprisingly the material that presented most problems for collection from the Internet was spoken data of all kinds. In particular, transcriptions of seminars and lectures were few and far between and could only be collected in a piecemeal fashion with a high ratio of human input for the material added. For this material therefore we had to rely overwhelmingly on existing corpora as described below. The situation was somewhat better for extracurricular texts in that we were able to download transcriptions of radio and TV broadcasts, made available from station websites. Again this could be annotated with time of broadcast and region in a semi-automatic way, although it proved difficult to achieve a representative regional spread with UK sources being particular scarce here.

3.2 Longman textbooks and Longman Spoken American Corpus

In addition to the web sources for textbooks, we were able to add to the corpus material from textbooks recently published by Pearson Longman. This was particularly useful in filling in some subject areas that were under-represented in the material initially collected. Another advantage of this source was that the textbooks were published very recently, mostly in 2008 and 2009, thus making the spread of years for textbooks in the corpus nearly match that of articles and extracurricular writing.

The textbooks selected for inclusion in the corpus covered 21 different academic disciplines, e.g., culture studies, law, computer science. The number of textbooks from each discipline to be included was limited to secure an even spread of data across the corpus. These textbooks were processed from PDF files in the same way as the web material.

A smaller additional source, though very welcome in view of the general dearth of this kind of material available on the Web, was from the Longman Spoken American Corpus. This was collected in 1995 and represented a range of academic-related scenarios.

3.3 Existing corpora with academic content

Material was also taken from the academic sections of the British National Corpus. This comprised 56 articles from 13 different academic disciplines (e.g., literature, art, chemistry) published between 1975 and 1993. Attention was paid to the current relevance of the material in order to secure the up-to-date character of PICAE. Academic data from the ANC (written and spoken) that were already part of the Longman Corpus Network were also included. This included six textbooks of academic disciplines such as architecture and education as well as spoken academic data.

4. Encoding

Leaving aside texts taken from existing corpora, all newly-collected documents in the corpus were converted from either PDF or HTML format and cleaned to reduce unwanted material. The texts were then tokenised, lemmatised and tagged, in preparation for being viewed in a corpus query tool.

4.1 Text cleaning

An initial element in this task was that of resolving some issues of character encoding arising from the conversion from texts from PDF and HTML format. This involved running the text through a script to rewrite various problematic characters such as ligatures and other unrecognised symbols into a standard format. Furthermore, in conversion from HTML format, text headings were often corrupted and so a script was added to recognise and remove these.

The further task of removing unwanted material from the documents begs the question of what material should be taken to be an integral part of the text and what was superfluous. One standard issue with PDF conversions was the presence of header information such as title and page number on every page. Retaining this information in the text would not only lead to disruption in the text, but would also skew the corpus statistics for words typically occurring in header information. As converted PDF files fell largely into one of a small number of header patterns, it was possible to write scripts to automatically remove this information. Using these scripts it also proved possible in the majority of cases to rejoin paragraphs that were separated by the header information. In general, paragraphs in the converted PDF documents are converted into single lines, allowing paragraph breaks to be easily added to the corpus.

Other issues of text inclusion or removal were not so clear cut, however. For example, repetitions or near repetitions of text would typically occur in the front or back material in each issue of some journal or university factsheet. It could be argued that this material is part of the text itself, but at the same time it can have an effect on the perceived representativeness of the corpus. The policy in this situation was driven mainly by practicality - only where such sequences were noticed as occurring were attempts made to filter them out.

One problem associated in particular with scientific texts, which formed a significant proportion of the curricular writing, was the occurrence of mathematical formulae, diagrams and tables etc. As far as was practicable these were removed from the text.

4.2 Lemmatisation and Tagging

Each document in the corpus was provided with initial header information, detailing values for the various corpus attributes. The cleaned file was then tokenised by dividing according to spaces in the text (as noted above, the files were already provided with markers of paragraph breaks). Sentence breaks were added by the tagger/lemmatiser. This process was applied also to the texts taken from the pre-existing corpora, i.e. the tagging supplied in the corpus was removed and the file converted back to raw text. This was done to allow for uniformity in word-division and tagging.

The corpus was then lemmatised and tagged in one operation by the English TreeTagger, under licence from the University of Stuttgart, Germany.

5. Composition

An overview of the composition of PICAE is given in the following tables and charts. Table 1 shows the number of words of each component, i.e. *Written, Written Curricular, Written Extracurricular, Spoken, Spoken Curricular, Spoken Extracurricular,* as well as the number of words of each text type.

Table 1: Components of PICAE

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*
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**

With regard to the date of publication Figure 2 shows that almost two-thirds of the material was published between 2000 and 2009. This is due to the World Wide Web being the main source and the inclusion of recently published Longman textbooks. The materials published in the 1970s and 80s (5.4%) were mainly sourced from the British National Corpus. Each file was assessed for its relevance before including it in PICAE.

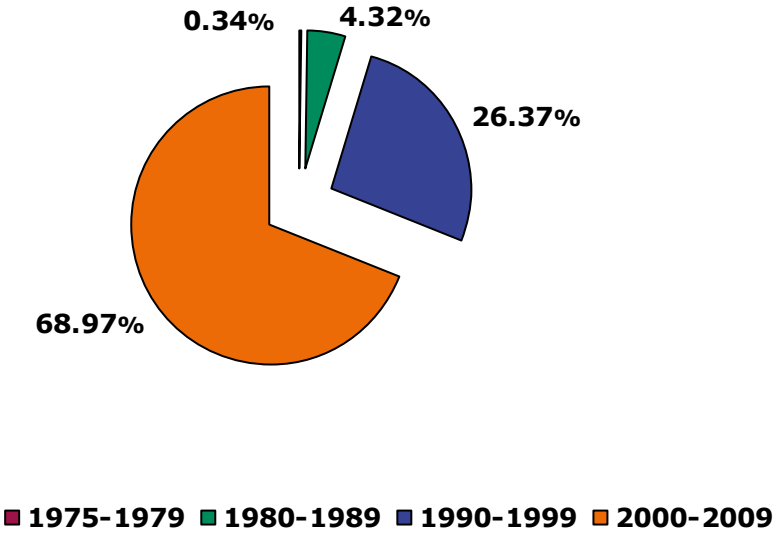

Figure 2: Composition by decade of publication

As mentioned above PICAE contains material from five English varieties. Figure 3 shows that 33% of the material could be classified in terms of its English variety. When publications had multiple authors or were published by international publishers, they were subsumed under 'global'.

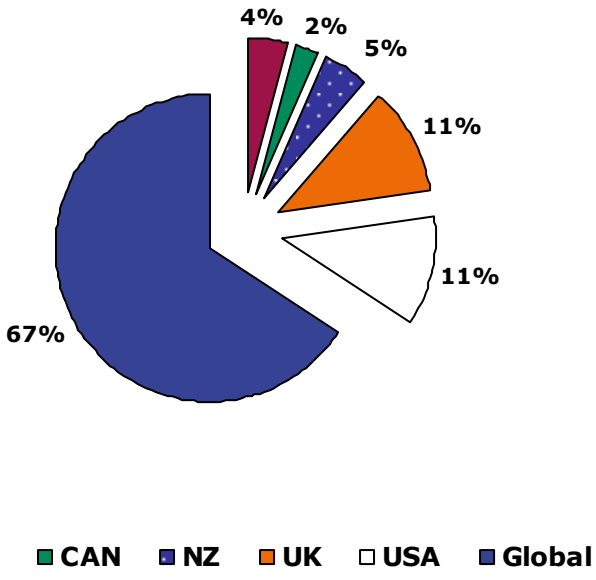

Figure 3: English varieties represented in PICAE

Figure 4 provides an overview of the text types PICAE comprises and their percentage. Textbooks form the largest category followed by journal articles and university/student/alumni magazines. Text types such as seminars and lectures make up only 3% due to the limited availability of spoken curricular data.

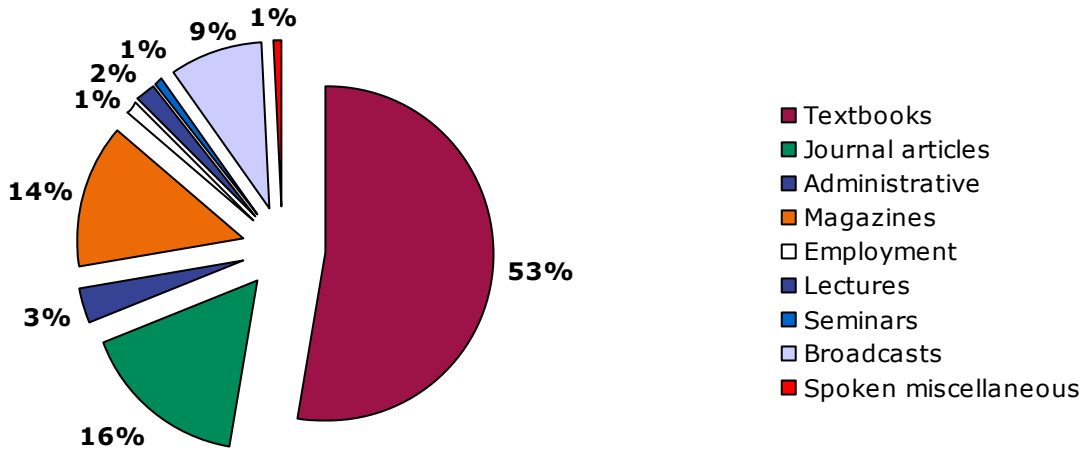

Figure 4: Text types represented in PICAE

The final figure shows the source corpora used in the compilation of PICAE as percentage. As mentioned above the World Wide Web was the primary source providing over half of the material used in PICAE. The second largest corpus consists of chapters form textbook published by Longman and accounts for one third of the corpus material.

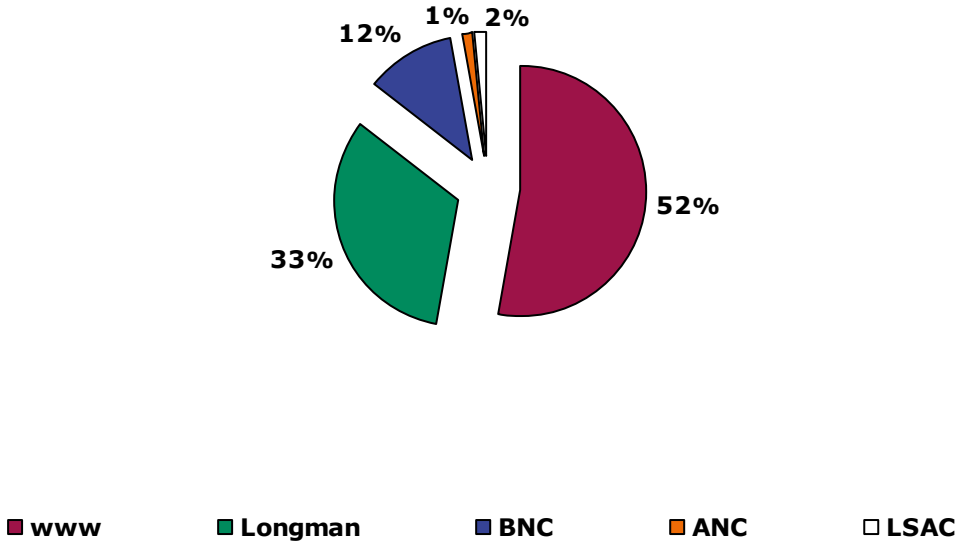

Figure 5: Source corpora used in PICAE

Table 2 shows the categorisation of the written curricular component of PICAE. Each of the four fields of study contains materials from seven academic disciplines to ensure that the corpus is representative of the academic register. The number of words per academic discipline as well as the total number of words per field of study and its percentage are provided. The four fields of study are currently not equally balanced. Professions and applied sciences are slightly overrepresented whereas the Humanities are slightly underrepresented.

Table 2: Fields of study and academic disciplines represented in PICAE

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 humanities	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 21%	**Total**	6,463,510 25%	**Total**	6,026,100 23%	**Total**	8,243,572 31%

6. Corpus-based research: The Academic Collocation List

The Academic Collocation List (ACL) should serve as an example for how PICAE is currently put to use. In 2010 Pearson decided to develop a list of the most frequent and pedagogically relevant collocations in written academic English discourse in order to produce lexicographic resources to facilitate, for example, EAP material development, item development, and validity research into PTE Academic. This list is derived from the written curricular component of PICAE, which comprises over 25 million words from 333 documents covering 28 academic disciplines as listed in Table 2.

The academic collocation list was developed in three stages: (1) computational analysis of the written curricular component of PICAE; (2) manual vetting based on target part-of-speech combinations and quantitative parameters, i.e. normed frequency per million ≥1; normed frequency per million in all four fields of study ≥.2; Mutual Information score ≥3; T-score ≥4 and (3) expert review to judge whether each collocation is pedagogically relevant.

Table 3 contains sample entries from the academic collocation list together with the following quantitative parameters for each collocation:

Table 3: Academic Collocation List: Sample entries

			Normed frequency in PICAE						
Pre-Collocate	**Academic Word**	**Post-Collocate**	**per million**	**Applied Sciences**	**Huma-nities**	**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

As the corpus is available for research, please contact pltsupport@pearson.com with any queries relating to using PICAE in corpus-based research.

References

Bailey, A.L. (2009). Defining Academic English: An Analysis of Practitioners' Perspectives. Paper presented at the Council of Chief State School Officers Assessment Conference, Los Angeles, CA. Retrieved June 30, 2010 from www.ccsso.org/content/PDFs/NCSA09_133_Bailey.pdf

Gillett, A. & Wray, L. (2006). EAP and Success. In Gillett, A. & Wray, L. (eds.) *Assessing the Effectiveness of EAP Programmes.* Papers from BALEAP PIM, University of Herfordshire, UK. Retrieved April 20, 2010 from http://www.uefap.com/articles/aeeapp.pdf

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 *Proc. Euralex.* Lorient, France, July: 105-116.