

Research Summary: The Pearson International Corpus of Academic English (PICAE)

Kirsten Ackermann
Pearson, London, UK
Kirsten.Ackermann@pearson.com

John H.A.L. De Jong
Pearson, London, UK
John.dejong@pearson.com

Adam Kilgarriff
Lexical Computing Ltd.

David Tugwell
Lexical Computing Ltd

September 2010

1. Introduction

As part of the development programme for Pearson Test of English Academic it was decided in 2007 to compile a new 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 account of the English that students will need to understand and produce in order to be successful in academic settings where English is the language of instruction.

Pearson contracted Lexical Computing Ltd., a company with specialist expertise in the creation and exploitation of corpora. A proposal for a new academic corpus, larger than existing ones, was jointly developed. This summary describes the three phases to the creation of PICAE: design, collection and encoding and provides an overview of the current composition of the corpus.

2. Design

PICAE comprises over 37 million words including 13% spoken and 87% written material. As PICAE was designed with reference to the question, what English does a non-native speaker need in order to be successful in academic settings where English is the main language, the corpus was to include both the English needed for academic work (72%) (hereafter referred to as curricular English) and the English needed for various aspects of extracurricular life, e.g., dealing with university administration, reading student magazines (28%) (hereafter referred to as extracurricular English). PICAE covers American, Australian, British, Canadian and New Zealand English.

The curricular material includes a wide range of academic subjects covering the four main academic disciplines, namely humanities, social science, natural & formal science and professions & applied sciences. It also comprises lectures, seminars, textbooks and journal articles at undergraduate as well as postgraduate levels. The extracurricular material 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 design.

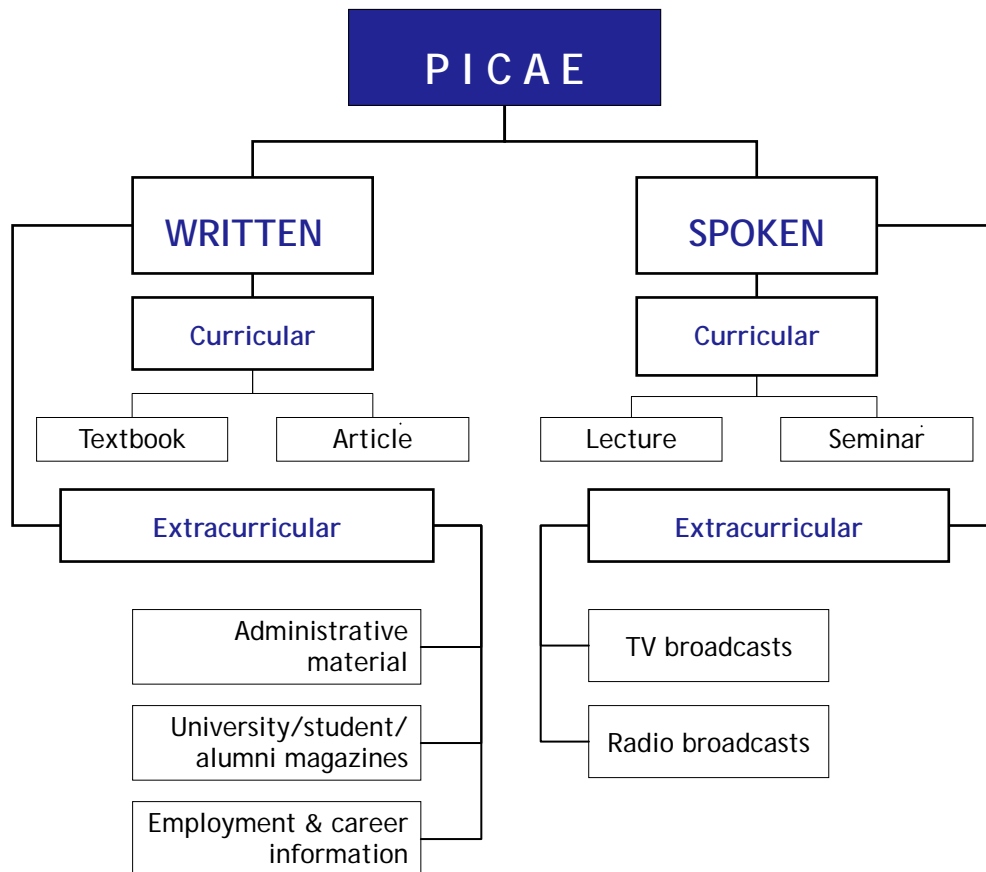


Figure 1: Design of PICA E

3. Collection

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

- 19.6 million words from the World Wide Web
- 12.1 million words from the 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 of academic English from the American National Corpus

The principal source was the World Wide Web. In addition material from recently published Pearson Longman textbooks were added to the corpus. This was particularly useful for achieving an equitable representation of academic disciplines.

The Longman 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 was limited to secure an even spread of data across the corpus. 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 journal articles and extracurricular writing.

Spoken material from the Longman Spoken American Corpus was included in the corpus. This material represents a range of academic-related scenarios and was originally collected in 1995.

Material was also taken from the academic sections of the British National Corpus. This comprises 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 American National Corpus that were already part of the Longman Corpus Network were also added to the corpus. This included six textbooks of academic disciplines such as architecture and education as well as spoken academic data.

4. Encoding

4.1 Text cleaning

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 the Sketch Engine, a corpus query tool.

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, e.g., ligatures into a standard format. Another 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 have led to disruption in the text, but would have also skewed the corpus statistics for words typically occurring in header information. Consequently, this information was automatically removed.

Repetitions or near repetitions of text would typically occur in the front or back material in, e.g., student magazines. It could be argued that this material is part of the text itself, but at the same time it could have a negative effect on the perceived representativeness of the corpus. Therefore where such sequences were noticed they were filtered 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 Lemmatization and tagging

Each document in the corpus was provided with initial header information, detailing values for the various corpus attributes which include document title, language mode, category, academic subject, extracurricular field, region and year of publication as well as source.

The cleaned file was then tokenised by dividing according to spaces in the text. This process was also applied 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 whole corpus was then lemmatised and tagged in one operation by the English TreeTagger¹, under licence from the University of Stuttgart, Germany.

¹ For more information go to <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>

5. Composition

An overview of the composition of PICAЕ is given in the following tables and charts. Table 1 shows the number of words of each component and its percentage.

Table 1: Components of PICAЕ

Component of PICAЕ	Number of words in million	Percentage of each component in PICAЕ in %
WRITTEN	32.5	87
Written Curricular	25.6	69
Textbooks	19.6	53
Journal articles	6.0	16
Written Extracurricular	6.7	18
Administrative	2.2	6
Magazines	4.3	12
Employment	0.4	1
SPOKEN	4.6	13
Spoken Curricular	1.1	3
Lectures	0.8	2
Seminars	0.3	1
Spoken Extracurricular	3.3	9
Broadcasts	3.3	9

Table 2 shows the four fields of study that were used to categorize the academic disciplines represented in PICAЕ. This categorization largely follows the list of academic disciplines published by Wikipedia.

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

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

In regard to the date of publication the following chart shows that almost two-thirds of the material was published in the last decade (2000-2009). This is due to the 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.

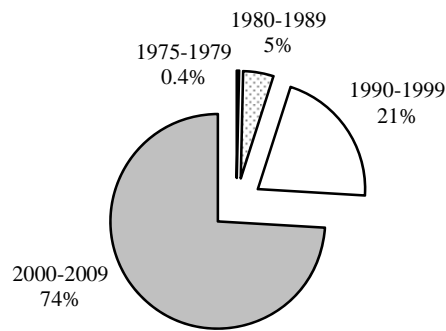


Figure 2: Composition by date of publication

As mentioned above PICAE contains material from five English varieties. Figure 3 shows that 30% 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'.

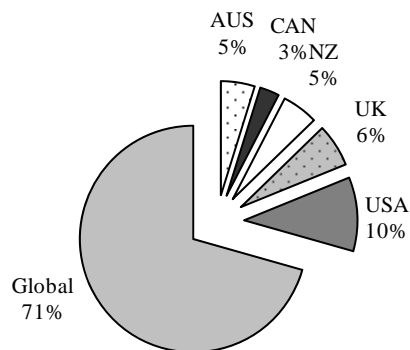


Figure 3: English varieties represented in PICAE

6. Conclusion

As English has become the lingua franca of today's globalized world and international student mobility is constantly increasing, proficiency in academic English is essential for students who want to succeed in an English-speaking academic environment.

PICAE was built to be representative of the English students encounter in campus life both inside and outside their studies. The corpus provides a source of empirical information on frequency and range within the register of academic English. PICAE comprises data from existing corpora and from the World Wide Web. The approach allows the collection of a great variety of texts in written as well as spoken academic English. The web provides up-to-date language data, so that 74% of the data originate from the last decade; 55% of the corpus data are from 2008-2009.

Updates of the corpus on a three-year cycle will allow us both to follow linguistic change, and to balance the various components of PICAE. Over the months to come, we are planning a number of studies based on PICAE to further explore academic English, academic word lists and the many ways in which corpora can support language teaching and assessment.

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

If you are interested in using PICAE in corpus-based research, please contact pltsupport@pearson.com