EAP Materials: Insights from Corpus-Based Studies
Eleni Agathopoulou
Aristotle University of Thessaloniki

Abstract
Corpus-based research can offer valuable insights to teachers of English for Academic Purposes (EAP). Such research may demonstrate typical and less typical patterns of language use and highlight important phenomena commonly not included in conventional EAP syllabuses. Moreover, studies exploring non-native English corpora inform us on learner problems in academic English. To make prominent the pedagogical implications of corpus research for EAP, in the current paper I focus on corpus findings in three areas, namely, compound words, discourse-organising nouns and titles of scientific articles and offer suggestions for the pedagogical exploitation of these findings.
Key words: academic English, corpus research, teaching materials, compound words, discourse-organising nouns, titles of scientific articles

1. Introduction

As well known, teachers of English for Academic Purposes (EAP)\(^1\) often have to design their own courses and to create teaching materials or to supplement existing ones. This endeavour presupposes, among other things\(^2\), awareness of the essential features of the texts relevant with the students’ needs and of the problems the students may have in handling these textual features.

In recent years, knowledge of genre-specific features has been aided significantly by research employing electronic corpora and software tools for corpus search. Studies in this methodological domain have yielded informative findings regarding features of academic English, indicatively: rhetorical structure (e.g. Halleck & Connor 2006; Peacock 2002; Pho 2010) grammatico-rhetorical relations (e.g. Pho 2008; Salager-Meyer 1989), grammatico-functional descriptions (e.g. Ferguson 2001), lexicogrammatical patterns (e.g. Hunston & Francis 1999), cohesion (e.g. Bondi 2004, Gray 2009 and references there; Römer & Wulff 2010; Swales 2005), metadiscourse (Harwood 2003; Hyland 2005; Samson 2004), hedging (e.g. Hyland 1998a,b; Poos & Simpson 2002), collocations (e.g. Curado Fuentes 2001; Lúzon 2000; Ward 2007), lexical bundles (e.g. Biber & Barbieri 2007; Csomay & Cortes 2010; Hyland 2008; Wood 2010), idioms (e.g. Simpson & Mendis 2003) and word-frequencies (e.g. Hyland & Tse 2007; Martínez et al. 2009). Corpus-informed studies have also made prominent cross-disciplinary differences in particular academic genres (e.g. Bruce 2010; Conrad 2001; Golebiowski 2009; Harwood 2005a; Hyland 2001, 2009; Samraj 2005; Thompson 2009).

Even more recently, significant insights regarding students’ problems in handling English academic genres come from studies drawing on learner corpora usually alongside comparable corpora from native peers or/and from expert native writers (Granger 2003,

\(^{1}\) In this paper EAP will be used as an umbrella term that includes English for General Academic Purposes (EAGP) and/or English for Specific Academic Purposes (ESAP).

\(^{2}\) For instance, knowledge about language learning and teaching theories, principles of course design and needs analysis models and methods.
10

2004; for more recent literature reviews, see Guilquin et al. 2007 and Granger 2009). This research has disclosed discrepancies between English native and non-native use regarding, for example, the expression of: author stance/engagement (e.g. Charles 2009; Hasselgärd 2009; Hatzitheodorou & Mattheoudakis 2007, 2008, 2009; Hyland 2009; Neff et al. 2003), hedging (Hinkel 2005), cohesion (e.g. Condrad 1999; Gray 2009; Hinkel 2003), modality (Ajmer 2002; Hyland & Milton 1997), aspect (Leńko-Szymańska 2007), tense, aspect and voice (Hinkel 2004), use of first person pronoun (Martínez 2005) and collocations (Nesselhauf 2005, 2007).

Ideally, EAP teachers should be able to do their own research based on corpora relevant with their students’ needs. However, this is may not be feasible due to lack of corpus-searching skills or access to specialized corpora. Hence, the next best thing for EAP teachers is to be aware of corpus-based research, as this research may (a) “highlight important aspects of a text or text collection that may go unnoticed otherwise” (Römer & Wulff 2010: 101), (b) reveal that language descriptions in commercial EAP textbooks are misleading (Barbieri & Eckhardt 2007; Harwood 2005b,c; Hyland 1994; Paltridge 2002; Swales 2002; Sznajder 2010; WonHo Yoo 2009) or underrepresent aspects of academic language use (e.g. Chen 2010; Pho 2008) and (c) confirm or disprove intuitive assumptions regarding learners’ language problems (e.g. Rundell & Granger 2007).

While many authors of corpus studies underscore that their findings have important implications for EAP teaching, relatively few of these authors suggest specific classroom activities. The latter point is not meant as a criticism, but simply to underline that incorporating the findings of corpus research is often left to the ability and the imagination of EAP teachers (cf. Conrad 1999: 16). This takes us to the aim of the present brief paper: to bring into focus some insights from corpus research and suggest how these insights can be implemented for the EAP classroom, as elaborated next in Section 2. The paper is rounded off by Section 3 which contains the recommendation of some corpus-based studies that demonstrate how their findings can be translated into teaching practice.

2. Some corpus-based findings and suggestions for classroom implementation

This section is divided into three subsections, each one focusing on a specific area of interest for EAP, namely: compound words (3.1), discourse-organising nouns (3.2) and titles to research articles (3.3).

2.1. Compound words

English compound words (hereafter ‘compounds’) consist of two or more words or stems (see, e.g. Bauer 1978: 49), for example: drug research, borderline personality disorder, cannabinoid-1 receptor blocker, filter-feeding bivalve molluscs, micro-column liquid chromatography-tandem mass spectrometry. Corpus research has corroborated the importance of compounds for EAP. Salager’s (1984) small corpus analysis revealed that compounds are considerably more frequent in medical than in general English texts (0.87% vs. 9.76% respectively) and in Biber and Gray’s (2010) large corpus analysis compounds occurred about three times more in English research articles than in oral discourse. Even more interesting for teachers of English for Specific Academic Purposes
may be Horsella and Perez’s (1991) corpus study of chemical English. This study disclosed that the frequency of compounds varied from 12.75% in introductory course texts to 37.17% in more advanced course texts and, impressively, to 50.08% in texts of graduate level and those concerning communication among scientists; also, the higher the text level, the larger was the number of multi-word compounds.

Halliday (1993: 77) points out that among other features that contribute to the high lexical density of scientific English, compounds may be “the hardest examples to process” and second language research has found that EAP learners have difficulty both with the structural form of such compounds and with the selection of proper contextual clues to interpret their meaning appropriately (Olshtain 1981; see also Trimble 1985). EAP/ESP textbooks and articles offer a variety of exercises on compounds, for example, “Form compounds from the following words” and “Find X two- or three-word compounds in the text” (Saslow & Mongillo 1985), “Turn these phrases into noun compounds” (Master 2003), “Choose the correct definition for the noun compound” (Master 2003; Williams 1985) and “Use a compound to name an entity” (e.g. A device that scans bar codes = a bar code scanner; Boeckner & Brown 1993)3. However, while exercise types that focus on the structure of compounds abound, to my knowledge, there is a scarcity of exercises/activities that may enhance learners’ ability to select the proper contextual clues for interpreting compounds. The activity I suggest below aims at the treatment of the latter type of learner problem.

First a slight diversion is in order to explain the theoretical background of this activity. According to the “linear on-going accumulation principle”, “each discourse element and its textual presentation at any stage of the ongoing process of discourse is premised upon the collection of all the foregoing elements and builds upon it” (Candlin & Lotfipour-Saeedi 1983, cited in Campoy & Carda 1998: 63). To illustrate how this is related with compounds, consider (1)-(5) appearing in this order in a text (excerpted from Halliday ([1999] 2004: 108).

(1) …the fracture of glass…
(2) …preexisting cracks . . .
(3) …Cracks in glass can grow…
(4) …the slow growth of cracks…
(5) …the crack growth …

The compound “crack growth” (5) is gradually built upon elements that appear first in a sentence (3) and then in a noun phrase + preposition + noun phrase construction (4). Also, in the beginning of the text there is a synonym for the first constituent noun of the compound (‘fracture’) which first appears in an adjective + noun phrase (2).

I suggest that the “linear on-going accumulation principle” can be exploited for pedagogical purposes. Learners can be asked to mark phrases or words in a text that relate to the eventual formation of a compound in this text, as demonstrated in the following text excerpts.

---

3 Learners may give left-headed instead of right-headed interpretations, that is, translate engine alterations as “a kind of engine”. For a treatment of this type of difficulty in large contexts see Williams (1985).
4 For a large variety of exercise types on compounds see Master (op. cit.).
For the first time, researchers have established how much exercise women should be able to do for their age and found that their capacity is slightly lower than men's. It also declines a bit faster than men's as they grow older.

Women whose exercise capacity was less than 85 percent of what it should be were twice as likely to die within eight years, the researchers found.

In 1930, the death rate from lung cancer among men was less than 5 per 100,000 population per year. By 1950 it had quintupled to more than 20; today it is above 70. The CDC in November 1993 said lung cancer death rates increased from …

In 1800, the majority of the world's population resided in Asia and Europe, with 65 percent in Asia alone. By 1900, Europe's share of world population had risen to 25 percent, fueled by the population increase that accompanied the Industrial Revolution. Some of this growth spilled over to the Americas, increasing their share of the world total. World population growth accelerated after World War II…

Additional merits of the activity described above are that it fosters awareness of cohesive relations in texts and, on a practical plane, it can be carried out using texts already included in the EAP materials.

2.2. Discourse-organising nouns

This section deals with a category of abstract nouns in noun phrases that refer to “a stretch of discourse as a linguistic act” and are thus termed “metalinguistic labels” or just “labels” (Francis 1994: 83). Some examples are: argument, assumption, belief, claim, comment, conclusion, concept, contention, controversy, debate, definition, dispute, example, excerpt, idea, illustration, notion, opinion, passage, point, position, principle, process, quotation, rationale, statement, theory, term and view. More general categories of abstract nouns with discourse-organizing function have been termed ‘shell nouns’ (Aktas & Cortes 2008 and references there) or ‘signalling nouns’ (Flowerdew 2003, 2006, 2010).

Phrases containing the discussed nouns may have anaphoric or cataphoric reference (‘retrospective’ or ‘advance’ labels, respectively, in Francis, ibid.), as demonstrated by the elements in bold type in (10) and (11) respectively.

---

5 Source: [http://health.dailynewscentral.com/content/view/1423/63](http://health.dailynewscentral.com/content/view/1423/63)


Particular concern is expressed regarding the field of R&D, because of the dominance of military expenditure in this area and the scarcity of qualified people. This point is made by Glyn and Harrison. . .(from Charles 2003: 317)

Spending on Medicare, the federal health program for the elderly and the disabled, grew 5.7 percent in 2003, to $283.1 billion, compared with an increase of 7.6 percent in 2002. Out-of-pocket payments were the only major source of health spending to increase faster in 2003 than in 2002, the administration said. Federal health officials offered two reasons: the number of people without health insurance is increasing, and employers who provide coverage are requiring workers to bear more of the costs.

Research employing learner corpora has confirmed a serious underuse of discourse-organising nouns in academic argumentative essays written by university students of upper-intermediate to advanced levels of proficiency in English. Both L1 Chinese students (Flowerdew 2010) and L1 Greek students (Mattheoudakis & Hatzitheodorou (in print)) have been found to use the discussed nouns about three times less than their native English peers. A characteristic example from the latter study is the noun argument, which occurred 253 and 19 times in the native and the non-native corpora, respectively. Mattheoudakis and Hatzitheodorou (in print) also found that while the native English writers often modified discourse-organising nouns with adjectives to express stance, for example, strong/weak argument or heated/fiery debate, the Greek learners did not. Other learner problems identified in this area by Flowerdew’s (2006) study, which was based on corpora similar to those in his 2010 study, regard wrong collocations, for example “draw an impact” or “discrimination to” (instead of “discrimination against”), or noun omission after a demonstrative (e.g. “these”), which resulted in unclear cross-sentence reference. Mattheoudakis and Hatzitheodorou (in print) attribute the learners’ underuse of discourse-organising nouns to lack of instruction in this area, probably due to that noun phrases containing the discussed nouns can not be easily classified into categories of connecting phrases for teaching purposes, unlike adverb phrases that can be categorized under labels such as “Summarizing” (e.g. all in all), “Making a concession” (e.g. at any rate), “Time reference” (e.g. afterwards) etc. (Longman English Grammar 1988: 326).

Explicit instruction within EAP courses may be able to help learners to develop effective use of the discussed phrases. To this end, teachers can provide learners with texts rich in discourse-organising nominal phrases and discuss the function of these phrases. For example, excerpt (12) serves as an illustration of both anaphoric and cataphoric functions of the discussed nominals (underlined and in bold, respectively).

Education in global citizenship, or world citizenship, therefore, requires familiarisation with and understanding of the network of relationships and

---

9 Flowerdew (2006, 2010) deals with “signalling nouns”, which is a category of abstract discourse-organizing nouns broader than that of “metalinguistic labels”. For the purposes of the present paper I focus on the overlap in findings between Flowerdew’s and Mattheoudakis and Hatzitheodorou’s studies.
10 This may explain the attested overuse of adverbials for text cohesion by learners; see Mattheoudakis and Hatzitheodorou (in print) and references there.
processes that pertain to a complex and multifarious world, inclusive of the local community to which citizens belong. By engaging in aspects of this ongoing debate, which are pertinent to the discourse for citizenship education and civic culture, the following argument will emerge. A world-embracing vision derived from… (Golmohamad, M. 2009)

I also suggest the following activities.
Activity A
Remove the sentences containing these nouns and ask students to find the place of these sentences in the text. To illustrate, this can be done with the sentence marked in bold type in (13).

(13) There is a mistaken belief that literacy development doesn't truly begin until a child starts formal instruction at an educational institution. **This is a misconception.** In fact, literacy development also involves listening and speaking abilities, which begin prior to elementary school. There are many things parents can do to aid in their baby's literacy development.\(^{11}\)

Activity B
In addition to the original sentence include other sentences with discourse organizing nouns and ask students to choose the sentence they think that best completes the text, as in (14).

(14) Patriotism does tend to cloud or distort civic judgment, and this can have dire consequences for those at the sharp end of unjust national institutions and policies. _______________. It is not necessarily the case that political judgments made by people who love their countries will be distorted and it may be that the danger can be somewhat mitigated by making patriots aware of it.

1. But, again, this argument is not decisive. [The original sentence]
2. But, again, this is not a sound argument.
3. But, again, this argument appears to be quite plausible.
4. But, again, this argument is a fallacious one.

Activity C
Draw learners’ attention to collocations with discourse-organizing nouns. For instance, a cursory search for collocates of ‘this/these argument(s)’ in Google Scholar\(^ {12}\) yielded a wealth of such collocates shown below.

\(^\text{11}\)\(\) Examples of verb collocates: adopt, clarify, counter, display, defend, dispute, elucidate, endorse, expound, make, meet, prove, question, support, take.

---

\(^{11}\) Source: Pear, 2005 <www.ehow.com/facts_6064535_literacy-development-babies.html>

\(^{12}\) Google Scholar (http://scholar.google.gr/) is a searching engine that indexes the full text of scholarly literature in a wide range of research areas. Note that the ‘Advanced Scholar Search’ allows the option of searching articles only in specific subject areas, such as Social Sciences, Biology etc.
Examples of noun phrase + of preposition phrase collocates: the basis of, (a/the) consideration of, the form of, the foundation of, the implication of, the line of, the premise of, the robustness of, the structure of (a/the) consideration of.

Examples of adjective collocates (in premodifying or postmodifying position):
accurate, baseless, cogent, commonsensical, consistent, decisive, explicit, feeble, firm, fragile, groundless, inconsistent, implicit, indecisive, persuasive, robust, sound, specious, straightforward, strong, unconvincing, unequivocal, unfounded, unreasonable, unsound, weak.

Using appropriate texts, teachers can ask learners to pick out collocations with frequently used discourse-organizing nouns and, furthermore, to brainstorm about synonymous collocations. Also, providing lists of such collocations may help learners improve their essays. Importantly, this activity can offer a rich supply of expressions of stance, as some of the verb collocates and many of the adjective collocates of ‘argument’ shown previously. For further focus on such collocates, learners can be asked whether the authors of a text agree or disagree with an argument, what they think of a specific claim, idea, opinion etc. Alternatively, learners can be directed to trusted websites for individual discovery of the collocations under discussion.

2.3. Titles of scientific articles

Given the importance of heading scientific articles appropriately, textbooks on writing scientific articles usually contain a section with advice on titles. Here are some examples of such textbook advice selected for reasons which will be clear later in this subsection.

“It [=the title] should include a verb, which should be in the present tense. Find an active verb, and aim for the grammatical structure Subject–Active verb–Object, i.e., “Viral Interleukin-6 Blocks Interferon Signaling”. (Katz 2009: 159)

“Hanging titles (=titles with a colon) … appear pedantic, often place the emphasis on a general term rather than a more significant term, necessitate punctuation, scramble indexes, and in general provide poor titles. …. titles phrased as questions … become unintelligible, and in my view “question” titles should not be used” (Day 1988: 20-21).”

“Try to avoid questions in the title, unless the issue remains unsettled or you came up with a clear answer”. (Alexandrov & Hennerici 2007: 257)

Let us next turn to relevant findings from two recent corpus studies.

Soler (2007) searched a corpus of 480 research article titles in the biological (medicine, biology and bio-chemistry) and social sciences (linguistics, psychology and anthropology). Titles were categorized into the following four types: Full-sentence type, e.g. “Learning induces a CDC2-related protein kinase”, Nominal group type, e.g. “Acute liver failure caused by diffuse hepatic melanoma infiltration”, Compound type, e.g. “Romanian nominalizations: case and aspectual structure” and Question type, e.g. “Does the Flynn effect affect IQ scores of students classified as LD?”. The corpus analysis demonstrated a preponderance of the nominal group (42%-38%) and a scarcity of the question type (0%-2%) in the titles of research articles across both of the disciplines examined. Moreover, full-sentence titles headed articles in the biological
sciences exclusively (38%) and compound-group titles were used much less in the biological than in the social sciences (10% and 38% respectively). Ball’s (2009) analysis of 20 million research article titles revealed that from 1966 to 2005 the use of question-type titles increased very little in physics (0.25%-0.5%) but significantly in life sciences (0.0025%-2.3%) and even more so in medicine (1%-5%). Given that titles with questions may more easily attract one’s attention, in Ball’s account, “The widespread use of the question mark clearly correlates with the increasing competition in science and research” (: 11). Also, Ball speculates that in the future there may be a further increase in question-type titles in medical articles.

EAP teachers should therefore be aware that evidence from corpus research may fly in the face of textbook advice recommending full-sentence titles and the avoidance of nominal-group and question-type titles in research articles. Also the corpus-attested differences across disciplines in this area offer another useful insight that can be incorporated in the EAP materials.

I suggest the following activity to enhance learner awareness concerning types of titles in research articles. Ask learners to
- find the most cited or the most downloaded articles in an academic journal of their specialism at online publisher sites,
- classify these articles into the four types described previously and
- discuss which types occur more frequently.

To offer a small illustration, in the beginning of October 2010 my own search of the ten most downloaded articles in the Journal of English for Academic Purposes (JEAP) and the American Journal of Medical Quality (AJMQ) yielded the results exhibited in Table 1.

Table 1. Types of titles in the ten most downloaded articles in JEAP & AJMQ (October 2010)

<table>
<thead>
<tr>
<th></th>
<th>JEAP</th>
<th>AJMQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full-sentence type</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Nominal group type</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Compound type</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Question type</td>
<td>-</td>
<td>2</td>
</tr>
</tbody>
</table>

Although the above results come from a negligible sample size, note that they generally comply with Soler’s (2007) findings regarding cross-disciplinary differences in this area and with the relatively frequent appearance of question-type titles in medical research articles attested in Ball’s study.

Last, in view of Alexandrov and Hennerici’s (op. cit.) recommendation, a further activity, probably more suitable for postgraduate students, would be to scan the abstract or the discussion section in order to find out whether the author(s) came up with a clear answer or whether the issue remained unsettled.

3. Final notes
The aim of this paper was to provide a short illustration of how EAP practitioners “…can benefit from being wise consumers of corpus-based research, even if they do not do the research themselves” (Conrad 1999: 16). For this purpose, I have offered examples of how corpus findings can be incorporated into EAP materials. A variety of other tasks based on data from research on corpora of academic English can be found in studies that are at the interface of corpus linguistics and language teaching. Indicatively, see Harwood (2003) on person markers and metadiscourse, Charles (2009) on the projection of stance, Hyland (2008), Flowerdew (2002), Johns (1995, 2008), Hüttner et al. (2009), Pang (2002), and Swales and Lindemann (2002) on genre awareness, Suganthi (2009) on the construction of academic identities in dissertations, Hyland (1996) and Hyland and Milton (1997) on hedging, and Curado Fuentes (2001) on collocations; also see Aranha (2009) and Feak and Swales (2010) on genre-based syllabuses and Harwood (2010) on academic citations. Examples of relevant studies that draw on corpora of spoken academic English are Jones and Schmitt (2010) on vocabulary and phrases, and Simpson and Mendis (2003) on idioms. Last, some of the books that offer suggestions for pedagogical materials consonant with corpus-informed data are those by Hyland (2006), Swales (2004), and Swales and Feak (2000).13

References


13 For an example of a genre-based course that drew on Swales & Feak’s book, see Aranha (2009).


Thompson, P. (2009). “Shared disciplinary norms and individual traits in the writing of British undergraduates”. In M. Gotti (ed) Commonality and Individuality in Academic Discourse Bern: Peter Lang, 53-82.


