An investigation of noun frequencies in cohesive nominal groups

Andrew Drummond, King's College London

Abstract

General purpose academic word lists, such as Coxhead's (2000) academic word list, are widely used in the teaching English for Academic purposes. However, word frequencies in some micro-level aspects of academic discourse are yet to be determined, such as subject-specific word lists in some areas. This study has generated knowledge of noun frequencies in sentence transitions containing anaphoric lexical references to the preceding sentence. Investigating a corpus of approximately 5.6 million words of academic texts from the Social Sciences and Humanities has led to a list of 71 nouns most frequently used in cohesive nominal groups in these areas. This list was compiled with Antconc (Anthony, 2014) by examining eight syntactic structures containing an anaphoric determiner and noun. The list can be used alongside more general purpose lists to support L2 academic writing development. As well as the main list, two significant sub-lists have been identified: a list of items particularly useful for anaphoric references to a citation and a group of nouns that nominalise processes. Four frequently occurring nouns in the data have been identified as forming partitive constructions with a cohesive aspect enabling the writer to narrow or broaden the range of analysis in the writing. In addition, there is a proposed order in which the eight cohesive structures investigated could be introduced within an EAP syllabus.

Key words: *cohesion in academic writing; lexical cohesion; lexical density; corpus linguistics; academic word lists*

Introduction

There are various approaches within the field of English for academic purposes (EAP) aimed at improving the language skills of non-native speakers of English entering English-medium tertiary education. Research from different branches of applied linguistics has created knowledge which has informed the practice of EAP. This knowledge has enabled EAP to move towards data-driven, rather than intuition-based goals. For example, knowledge of written academic genres (Swales, 1990) has provided EAP practitioners with guidelines on how to construct academic texts in the manner of a target discourse community. Other forms of discourse analysis, such as Halliday and Hassan's work on cohesion (1976) and Kaplan's work (1967) on contrastive rhetoric, have enabled EAP practitioners to describe more accurately features of texts beyond the level of the sentence. More recently, Coxhead (2000) used Corpus Linguistics methods to devise an Academic Word List (AWL) which provides a

basis for moving students from general vocabulary usage towards lexical competence in an academic register.

However, there remain aspects of EAP practice that are not yet sufficiently informed by data-driven research. Whilst Coxhead's AWL and subsequent word lists are able to state word frequencies across whole corpora of academic writing, certain patterns of discourse may evidence a more specific set of lexical items. The present study seeks to explore word frequencies in one such area. Lautamatti's topical structure analysis (1987) describes three types of sentence transitions in academic texts which contribute to coherence in texts. For example, a sequential progression occurs when the rheme of a sentence is recycled as the theme of the subsequent sentence. When this recycling occurs, it is often the case that a lexically denser, cohesive form of words is used in the second sentence, in order to avoid verbatim repetition. The lexis that is frequently used in these sentence transitions has generative power to produce lexical density and strong cohesive links in academic writing.

The present study is a corpus-based investigation into exactly what lexis frequently occurs in these cohesive phrases of this type. Its aim is to provide a list of nouns most frequently appearing in these phrases in a corpus of academic writing from the fields of Social Sciences and Humanities. With this data-driven knowledge, it is hoped that EAP practitioners will be better equipped to scaffold good practice in this area.

Literature Review

Writing differs from speech in the respect that: '[it] tends to be lexically dense, but grammatically simple; spoken language tends to be grammatically intricate, but lexically sparse' (Halliday, 1994, p. 61). The lexical density of a text is measured by comparing the proportion of content words to function words. According to Johansson (2009, p. 65), 'an academic text with a high proportion of content words … is able to provide more information than a non-academic text of similar length.' In order to write academic text fluently, attention needs to be paid to the construction of complex noun phrases. As Halliday, Matthiessen, and Matthiessen (2004, p. 655) state: 'The nominal group is the primary resource used by grammar for packing in lexical items at high density'. Accordingly, giving students of EAP the productive power to construct lexically dense nominal groups can be a significant factor in the process of learning to write according to academic conventions.

Halliday and Hassan (1976) identified a range of the cohesive features of texts allowing connections and references to be made beyond the level of the sentence. They categorise these as reference, conjunction, ellipsis, substitution and lexical cohesion. Halliday, Matthiessen, and Matthiessen (2004, p. 570) explain lexical cohesion as when: 'a speaker or writer creates cohesion in discourse … through the selection of [lexical] items that are related in some way to those that have gone before.' Lexical cohesion and lexical density overlap when these references, linking to previously occurring text, are also content words, as in the following example:

> "A calculation is then made whereby **operating costs are subtracted from turnover**. This **process** allows profit to be calculated ..."

Here, *process* is a content word in the sense that it is a noun which contributes substantive meaning to the sentence. It is also lexically dense as it replaces the entire phrase, *operating costs are subtracted from turnover* with a single word. It is cohesive in that its meaning can only be decoded by reference to the longer phrase in the previous sentence, 'instead of being interpreted semantically in their own right' (Halliday and Hassan, 1976, p. 31).

Some of the words that can deployed as references to link text are demonstratives including 'this', 'that', 'these', 'those' etc. Whilst lexical cohesion and reference are listed as separate categories of cohesion in their work, they also acknowledge that, 'there are many instances of cohesive forms which lie on the borderline between two types and could be interpreted as one or the other' (Halliday and Hassan, 1976:85). Nominal groups containing both a demonstrative reference word and a lexically cohesive noun are an example of a hybrid form that sits on the border of two Hallidayan categories. Here is an example of such a hybrid reference taken from the British National Corpus (BNC) (text J0V):

> "Over the last ten years the use of computers in the work of historians has increased dramatically. **This development** is undoubtedly due to the greater accessibility that personal computers have brought to computing …".

As in the previous example, the noun phrase 'this development' recasts the entire, previously stated idea into a denser form of words as a form of substitution employed to avoid redundancy. For the purposes of this study, I refer to these kinds of word combinations, containing a reference word and a noun used for lexical cohesion, as

cohesive nominal groups. There are a number of reference words that operate in cohesive nominal groups to form the reference component. The following examples of cohesive nominal groups illustrate some of the variety of reference words possible in such structures, along with other syntactic variations such as the inclusion of adjectives in the nominal group. They are taken from texts within the British National Corpus (BNC); the specific text in which they appear is noted with the original BNC reference in brackets (BNC, 2007):

Table 1: *Examples of different syntactic structures*

'**This method** of reasoning identifies the conditions...' (EB2)
'This new **method** of composition can be seen…' (GUJ)
'Both **these methods** of creating a mortgage give…' (ABP)
'…an evaluation of **these** matrix **methods…'** (GUC)
'They viewed **such methods** as a necessary evil…' (G04)
'…**such** unconventional **methods**…' (CLN)
'**Such** a **method** was tested by Whitley…' (FNR)
'**Such** a circular **method** of approach…' (CAW)

Along with the reference word, the noun in the above phrases forms a surface level tie which can only be decoded with reference to earlier text. In the above examples 'method' substitutes for an entire, previously stated process in a lexically dense manner. As Thompson (cited in Mueller, 2015, p. 23) states: 'After a process has been introduced in a text, it can be encapsulated as a thing and be used as the basis for the next point in the text or become a participant in another process in the text'. When a process is encapsulated as a thing, the resulting noun is known as a nominalisation. Fang, Schleppegrell, & Cox (2006, p. 254) describe the process of nominalisation as follows: 'Nominalization enables something that has been presented in a series of clauses to be distilled into one nominal element. Such distillation enables a chain of reasoning to be developed by the writer.' Nominalisation, then, contributes to lexical density through the construction of lexically dense content words which define processes and stand in for longer stretches of text. These content words are often deployed with a lexically cohesive function within the text.

EAP students need to be able to both *decode* lexically dense references when reading and *produce* them when writing. As Fang, Schleppegrell, & Cox (2006, p. 254) state: 'Being able to recognize referential links is crucial to comprehending academic texts, and adopting this type of reference in writing is crucial to constructing clear and coherent texts.' Hylands and Tse (2007, p. 243) also suggest that 'novice users' may need help to decode nominalisations in longer stretches of complex language. In addition, there is evidence that some L2 undergraduates' writing needs development in the area of lexical cohesion. O'Keeffe (2000) shows that repetition of nouns is used disproportionately often as a cohesion strategy in undergraduates' writing. In addition, Sadighi (2012) found errors involving lexical cohesion to be the second most common in her analysis of cohesion in Iranian ESL students' writing. Drummond (2015) noted that [this + noun] structures in IsiZulu speakers' academic writing were most often repetitions of earlier nouns and were rarely deployed as a means of contributing to lexical cohesion and lexical density simultaneously, as in the above examples from the BNC.

Coxhead's AWL (2000) aims to provide the EAP field with a list of words occurring frequently in a wide range of academic texts in order to help with 'making principled decisions about which words are worth focusing on during valuable class and independent study time' (2000, p. 213). Similarly, Nation (2004, p. 3) has explained that the making of word lists, 'in the field of L2 learning and teaching is usually done for the purpose of designing syllabuses'. Word lists can also inform design choices made by EAP materials' developers (Coxhead, 2000, p. 214) in terms of what content to include and the type of activities chosen. Words lists such as the AWL are created by assembling a large body of texts, known as a corpus, and searching this material with software in order to find data on word frequencies. As a general purpose list, however, the AWL is not designed to inform which nouns are most frequently used in cohesive nominal groups. The development of a list of high frequency nouns operating in these phrases could provide a basis for 'principled decisions' on how the matter of lexical cohesion and lexically dense sentence transitions could be dealt with on EAP programs.

Coxhead's AWL (2000) has been influential in the field of EAP, forming the basis of professionally published EAP materials (Schmitt & Schmitt, 2005) as well as the content of EAP focused websites. However, Hyland and Tse (2007) have questioned whether the AWL is relevant across of all the fields that contributed to Coxhead's combined corpora: arts, commerce, law and science. The fields that this study is concerned with are narrower than

Coxhead's; it deals only with professionally published academic texts from the Humanities and Social Sciences. The rationale here is that academic language instruction at the institutional level is often divided between these areas and hard sciences. Although narrower than Coxhead's field of study, Humanities and Social Sciences still represent a large number of disciplines, and so knowledge gained by this study may still be applicable to a wide area.

Of course, the question of what the most appropriate breadth of spectrum is when constructing a corpus from which language data will be derived remains difficult to answer. There are many academic disciplines and distinct varieties of English. Is it even linguistically accurate to describe a single academic discipline, such as history, as a unified language field? Does political history share the same language conventions as social history? Whatever the case, for L2 students entering tertiary education without a clear specialisation and where EAP input is divided between hard sciences and 'everything else', the data this study generates will be keenly relevant. However, this investigation is to be undertaken with an awareness that constructing a broad spectrum word list may inadvertently conceal data on lexis that is, for example, highly relevant in Social Sciences but not in Humanities. If this kind of compromise is necessary, it shall be noted.

The purpose of this study is principally to provide a list of nouns which are found to occur frequently in cohesive nominal groups in the two fields mentioned. This list could then be used as a means of developing lexical density in the writing EAP students and improving receptive recognition of these structures in extended academic texts. The list might prove to be useful both as a means of reference for EAP students whilst writing, and as a reference for professionals devising syllabuses, materials and lesson plans. As Cobb has said (as cited in Byrd & Coxhead, 2010, p. 51): 'Learners like word lists, so let's give them good ones.'

Key Research questions

1. Which nouns are most frequently used in cohesive nominal groups for the purposes of achieving lexical cohesion and lexical density in published academic work across the Social Sciences and Humanities fields?
2. Which of the eight syntactic structures investigated (see below) appear most frequently in academic writing in these corpora?
3. What proportion of the most frequently occurring nouns in these structures appear in the General Service List (West and West, 1953) and AWL respectively?

4. How could knowledge generated by this study inform EAP reading and writing skills input?

Research methods

The corpora

The approximately 5.6m tokens making up the combined corpora for this study are derived from texts taken from the BNC. The BNC was first published in 1995 and totals around 100 million tokens of written and spoken British English. None of the academic texts are dated before 1975. The combined corpora used for the present study comprises selections from the Social Sciences and Humanities areas of the BNC in order to generate data useful to EAP students heading into these faculty areas.

Ideally, a corpus of language being used to inform EAP pedagogy would be entirely contemporaneous with the point at which the pedagogy was deployed. In practice, however, a text is a historical document as soon as it is written. One reason for this is that language use is not static but changes over time. For example, measuring occurrences of the word 'notwithstanding' in the Google Books corpus (Ngram Viewer, 2016) reveals a marked decline in use over the past 200 years, while 'furthermore' shows a slight increase. The BNC, then, containing text ranging between 20-40 years in age represents language use that will not be entirely the same today. However, word lists using BNC material, such as Nation's BNC/COCA lists (Nation, 2016:138) have recently been updated as contemporary resources for pedagogy and research. Nation's BNC wordlists were used by Dang and Webb (2013) in a study measuring spoken academic vocabulary in favour of General Service List based lists (West, 1953) as they might better represent 'current vocabulary' (Dang and Webb, p. 53). This indicates an ongoing degree of relevance for the BNC in this area of research.

The Social Sciences and Humanities selections from the BNC used here are also freely searchable by students and practitioners on lextutor.ca (Cobb, 2015). Such direct accessibility may allow students, practitioners and/or researchers to examine the same corpora in order to verify and/or add to the knowledge generated here. For example, lextutor.ca could be used in class to investigate which adjectives mostly commonly appear in [this + adj + process] as a complement to the present study.

Another relevant factor here is the size of corpora: 5.6 million tokens make up the corpora used in this study. This is a large amount of data and can contribute to more reliable results

than that of a smaller sample. As Coxhead (2000, p. 216) states: 'A corpus designed for the study of academic vocabulary should be large enough to ensure a reasonable number of occurrences of academic words.'

Since both Humanities and Social Sciences are broad areas, these selections from the BNC cover a wide range academic disciplines in each field. This is in keeping with the aim to produce generalised knowledge across these fields; a narrow range of disciples would not be able to do this as effectively. Here are the key data relating to the combined corpora:

Table 2: *The combined corpora*

	BNC Humanities (Human)	BNC Social Sciences (Soc Sci)
Variety of English	British	British
Total number of texts	87	64
Range of text size	5,000 – 45,000	5,000 – 45,000
Total number of running words	3,346,833	2,260,406
Average text length	38,469	35,318
Subject / (no. of texts)	Archaeology (4) Art (5) Research methods in Humanities (2) Cultural theory (1) Film studies (1) History (28) Literary Criticism (16) Music Studies (5) Philosophy (13) Politics (11) Religious Studies (1)	Anthropology (3) Addiction studies (1) Child care policy (2) Crime and prison studies (5) Deaf Studies (1) Economics (1) Education studies (3) Geography (3) Public Health (1) Law (1) Linguistics (8) Population studies (1)

		Psychology (10) Research methods in Social Sciences (3) Sexual politics (1) Social policy (2) Sociology (18)

Procedure

Tom Cobb kindly sent me the source files for Humanities and Social Sciences corpora available on lextutor.ca. I divided these large files into their constituent texts and, using Tagant (Anthony, 2015), I added parts of speech (POS) tags to each text. This means that each word in texts now had a word class tag, such as 'noun' or 'verb' attached to it, allowing for the language to be sorted according to its grammatical categories. With these tagged files, I used Antconc (Anthony, 2014) to search the corpora for eight syntactic varieties of cohesive nominal groups in order to ascertain which nouns occurred most frequently in these structures for the lexical cohesion purposes. Here are the eight syntactic strings under investigation:

Table 3: *The eight structures investigated*

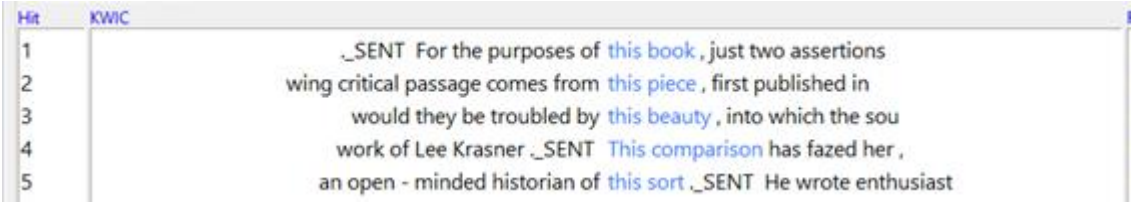
Syntactic form of cohesive nominal group	Search string used in Antconc	Example
This + noun	[this_DT *_NN]	This change…
This + adjective + noun	[this_DT * _JJ *_NN]	This major change…
These + noun	[these_DT *_NN*]	These changes…
These + adjective + noun	[these_DT *_JJ *_NN*]	These major changes…
Such + noun(s)	[such_JJ *_NN*]	Such chang(es)…
Such + adjective + noun	[such_JJ *_JJ *_NN*]	Such major changes…
Such + a/an + noun	[such_PDT a*_DT *_NN]	Such a change…
Such + a/an + adjective + noun	[such_PDT a*_DT *_JJ *_NN]	Such a major change…

These eight structures are not an exhaustive list of cohesive nominal groups but they are selected to represent some of the most frequently occurring ones in academic writing. In

these structures, the cohesive nominal group is headed by a determiner making an anaphoric reference: this, these and such. Noun phrases can be long and complex and it would be considerably difficult to investigate every possible structure making an anaphoric lexical reference. The variations shown above include determiners and nouns separated by one adjective, although, less frequently, there can be two. There are other determiners that can be used with a noun to produce a dense anaphoric reference, such as both, each and every, but do not have an exclusively anaphoric reference in the noun phrase. I have not included these in the study since the concordancing software is not able to distinguish their anaphoric uses. Similarly, I have not included instances of [the + noun] in this investigation as the concordance cannot differentiate between referential and non-referential uses of the definite article and a manual count of such phrases would have been impractical with a combined corpora of this size. Similarly, whilst the software, Tagant (Anthony, 2015), was successfully able to tag a range of demonstratives correctly, it did not identify 'that + noun' with sufficient accuracy when 'that' was a referential word. This is most likely due to the varied grammatical uses to which 'that' is put. Again, a manual count of cohesive noun phrases including 'that' would have taken a great deal of time and so this demonstrative was not included.

Each of the above searches generated a large number of concordance lines. For example, here are the first five concordance lines of a search for [This + noun] within the Humanities corpus:

Figure 1: *Example of concordance lines generated by [this + noun] search*



Hit	KWIC	
1	._SENT For the purposes of	this book , just two assertions
2	wing critical passage comes from	this piece , first published in
3	would they be troubled by	this beauty , into which the sou
4	work of Lee Krasner ._SENT	This comparison has fazed her ,
5	an open - minded historian of	this sort ._SENT He wrote enthusiast

The results generated by Antconc for the eight strings include occurrences of the nouns functioning as the head of a nominal group and as post-modifiers. I saved the results generated by these searches as text files and used this data to establish word frequencies for individual nouns with Antconc's word list tool for each of the two corpora respectively. For example, the noun 'process' appears 202 times in the Social Sciences corpus as part of a cohesive nominal group and 121 times in the Humanities corpus.

With the frequency data, I was able to construct a list of nouns commonly occurring in the cohesive nominal groups. In order to generate range data for the items on this list, I used the advanced search tool of Antconc to search for occurrences of these nouns in each of the eight strings at the same time. This advanced search was repeated for each corpus. I then noted the data from the concordance plot indicating the number of texts in which these structures appeared. For example, the noun 'process' appears as a part of a cohesive nominal group in 58 of the 64 texts in the social science corpus. These procedures produced frequency and range data for each corpus that enabled the construction of the word list below.

Results

This section details the key results from the above process. The following results provide a list of the most frequently occurring and wide ranging nouns in these structures with the inherent potential to contribute to lexical cohesion and lexical density. There is also a breakdown of the frequencies with which each of the eight structures investigated appeared in the corpus. In addition, I have recorded instances of nouns which featured significantly in one corpus but not the other. Finally, there is some data on how frequently cohesive nominal groups occur as the unmarked theme of a sentence.

The list of high-frequency nouns in cohesive nominal groups

The AWL is a much larger, general purpose list than this list of nouns in cohesive nominal groups (NICNGL). The specific purpose of the list below is to provide material to be used for the development of language awareness in the area of cohesive nominal groups. The time allotted to such an area within an EAP syllabus would be much smaller than that of general academic vocabulary and, as such, a smaller list seems more appropriate. A number of criteria had to be met in order to be selected for the list below. Words were included if they appeared at least 70 times in the eight cohesive structures listed above in the combined corpora, and more than 10 times per million tokens in each corpus. In addition, a range of at least 20% in both corpora was required for inclusion in the NICNGL. These criteria contribute to the construction of a list based on frequency and range data, and ensure that the list is considerably smaller than the AWL which seems appropriate since it is intended to inform a more limited area within academic discourse.

An additional criterion for inclusion is that a noun ought not to mostly be used to form a partitive structure, such as 'this sort of ….'. This kind of device is mentioned in a separate section below. Also, nouns which predominantly formed particular lexical phrases, such as

'in this way' were not included. In addition, the items comprising the NICNGL are intended to be available as lexical resources capable of referring to specific aforementioned text in an abstract rather than a concrete sense; the item 'man' was removed for not meeting this criterion. Similarly, nouns which deictically referred to the texts in which they appeared such as, 'book', 'chapter', 'paper' and 'section' were also removed on the grounds that they were direct references to concrete entities rather abstractions.

The list includes homonyms such as 'picture' which may be used in a concrete sense to refer to a painting in an art history text and metaphorically to refer to a description. No attempt to differentiate between senses of such items has been made here but, for EAP purposes, the concrete sense of the noun could be used as a point of departure to the abstract. Singular and plural forms of the nouns in this list have been added together to form the totals. The NICNGL is presented in Table 3 below:

Table 4: *Nouns in cohesive nominal groups list (NICNGL)*

Word	Total occurrences combined corpora	Total occurrences in Soc Sci	Soc Sci per million	Soc Sci range (no. of texts)	Soc Sci range %	Total occurrences in Human	Human per million	Human range (no. of texts)	Human range %	GSL	AWL
time	584	131	57.96	45	70.31	453	135	72	82.76	*	
case	541	304	134.51	52	81.25	237	70.63	66	75.86	*	
point	520	238	105.31	61	95.31	282	84.04	68	78.16	*	
view	401	220	97.35	48	75	181	53.94	55	63.22	*	
period	351	105	46.46	37	57.81	246	73.31	51	58.62		*
process	323	202	89.38	58	90.63	121	36.06	44	50.57		*
approach	313	221	97.79	43	67.19	92	27.42	40	45.98		*
question	304	162	71.68	45	70.31	142	42.32	50	57.47	*	
problem	299	190	84.07	51	79.69	109	32.48	50	57.47	*	
group	292	193	85.40	47	73.44	99	29.50	42	48.28	*	
area	277	152	67.26	48	75	125	37.25	51	58.62		*
change	275	168	74.34	45	70.31	107	31.89	51	58.62	*	
thing	243	79	34.96	28	43.75	164	48.87	59	67.82	*	
argument	239	119	52.65	40	62.50	120	35.76	41	47.13	*	
stage	235	138	61.06	39	60.94	97	28.91	40	45.98	*	
work	222	111	49.12	35	54.69	111	33.08	45	51.72	*	
issue	198	114	50.44	47	73.44	84	25.03	41	47.13		*
idea	195	96	42.48	43	67.19	99	29.50	39	44.83	*	
study	192	147	65.04	39	60.94	45	13.41	25	28.74	*	
situation	183	109	48.23	41	64.06	74	22.05	38	43.68	*	
sense	183	89	39.38	38	59.38	94	28.01	47	54.02	*	
difference	180	109	48.23	39	60.94	71	21.16	36	41.38	*	
Term	170	90	39.82	37	57.81	80	23.84	42	48.28	*	

theory	166	72	31.86	23	35.94	94	28.01	32	36.78		*
context	163	75	33.19	34	53.13	88	26.22	46	52.87		*
system	162	86	38.05	29	45.31	76	22.65	32	36.78	*	
example	155	82	36.28	34	53.13	73	21.75	37	42.53	*	
development	154	64	28.32	31	48.44	90	26.82	40	45.98	*	
figure	150	63	27.88	26	40.63	87	25.93	39	44.83	*	
factor	142	87	38.50	40	62.50	55	16.39	29	33.33		*
matter	140	43	19.03	23	35.94	97	28.91	42	48.28	*	
circumstance	133	76	33.63	40	62.50	57	16.99	31	35.63		*
century	129	77	34.07	28	43.75	52	15.50	28	32.18	*	
pattern	129	88	38.94	33	51.56	41	12.22	23	26.44	*	
distinction	126	53	23.45	26	40.63	73	21.75	36	41.38		*
assumption	125	74	32.74	33	51.56	51	15.20	27	31.03		*
claim	125	66	29.20	27	42.19	59	17.58	33	37.93	*	
account	123	61	26.99	25	39.06	62	18.48	32	36.78	*	
Fact	121	52	23.01	30	46.88	69	20.56	32	36.78	*	
position	121	64	28.32	28	43.75	57	16.99	37	42.53	*	
activity	118	76	33.63	30	46.88	42	12.52	22	25.29	*	
aspect	117	69	30.53	35	54.69	48	14.30	31	35.63		*
country	116	61	26.99	24	37.50	55	16.39	25	28.74	*	
method	113	60	26.55	26	40.63	53	15.79	29	33.33		*
principle	113	47	20.80	24	37.50	66	19.67	28	32.18		*
category	112	67	29.65	33	51.56	45	13.41	25	28.74		*
information	107	53	23.45	20	31.25	54	16.09	18	20.69	*	
relationship	106	68	30.09	29	45.31	38	11.32	28	32.18	*	
feature	104	62	27.43	33	51.56	42	12.52	23	26.44		*

concept	103	62	27.43	29	45.31	41	12.22	21	24.14		*
effect	103	63	27.88	28	43.75	40	11.92	30	34.48	*	
statement	103	32	14.16	23	35.94	71	21.16	35	40.23	*	
conclusion	101	50	22.12	26	40.63	51	15.20	30	34.48		*
event	101	37	16.37	24	37.50	64	19.07	39	44.83	*	
condition	98	46	20.35	27	42.19	52	15.50	27	31.03	*	
interpretation	97	51	22.57	23	35.94	46	13.71	28	32.18		*
attitude	97	49	21.68	27	42.19	48	14.30	25	28.74		*
word	96	27	11.95	17	26.56	69	20.56	35	40.23	*	
instance	92	43	19.03	43	67.19	49	14.60	30	34.48		*
evidence	91	34	15.04	23	35.94	57	16.99	27	31.03		*
level	91	50	22.12	29	45.31	41	12.22	26	29.89	*	
line	87	31	13.72	21	32.81	56	16.69	32	36.78	*	
state	86	32	14.16	17	26.56	54	16.09	30	34.48	*	
perspective	85	48	21.24	25	39.06	37	11.03	22	25.29		*
notion	82	35	15.49	19	29.69	47	14.01	25	28.74		*
structure	80	41	17.82	20	31.25	39	11.62	21	24.14		*
belief	79	34	14.78	21	32.81	45	13.41	19	21.84	*	
purpose	75	35	15.21	17	26.56	40	11.92	25	28.74	*	
practice	71	35	15.21	21	32.81	36	10.73	22	25.29	*	
occasion	71	26	11.30	13	20.31	45	13.41	30	34.48	*	
picture	71	25	10.87	17	26.56	46	13.71	28	32.18	*	

This list could be given students to aid with the development of lexical cohesion and lexical density in their writing. In addition it could inform EAP materials and syllabus design. For example, exercises could be devised in which processes were stated and were followed by a gapped sentence in which a rheme had become the theme. Students would then be required to choose appropriate nouns to make anaphoric, lexical reference back to the process. This is illustrated below:

> 'In order to promote critical thinking, students are trained in questioning the methodology presented in research articles and consciously adopting positions contrary to the writers' theses statements. This _____ is way of explicitly scaffolding procedures which later may become more automatic.'
> *Which of these nouns can complete the second sentence? A) practice B) notion C) method D) concept*

The total number of occurrences of nouns in the NICNGL in the combined corpora is 12050 which is over 33% of the total number of occurrences of the eight structures under investigation. This is a substantial proportion of the total number of occurrences of these structures in the combined corpora and indicates that familiarity with these items could provide considerable productive writing power to aiming to develop lexical density and cohesive links in their texts.

Proportion of items from the NICNGL also found in the GSL and the AWL

The purpose of the NICNGL is to enable EAP students to develop a particular feature of academic writing. It is does not necessarily follow, though, that the nouns most commonly used in cohesive nominal groups are predominantly 'academic words'. Here is a table showing the frequencies with which these 71 words appear in the GSL and AWL:

Table 5: *Percentage of NICNGL items found in the GSL and AWL*

Word list	Occurrences on the NICNGL	Percentage of NICNGL
GSL	46	65%
AWL	25	35%

According to Coxhead (1998), the general service list constitutes around 76% of the vocabulary of academic texts and the AWL approximately 10% of academic texts. With this in mind, the fact that 35% of the NICNGL is composed of items from the AWL indicates that these types of cohesive noun phrases contain a significantly higher proportion of academic vocabulary than in academic texts in general. In terms of pedagogy then, there is evidence that academic vocabulary should form a significant amount of the input given in this area, especially for those who are already familiar with discourse patterns involving cohesive nominal groups. Equally, for students unfamiliar with such structures, there is much relevant and familiar vocabulary from the GSL which might lessen the cognitive burden on those still acquiring the structures.

Using items for the NICNGL to evaluate citations

Further qualitative analysis of the NICNGL suggests that a subset of this list could be particularly useful for the purpose of evaluating citations in a sentence subsequent to cited material. Here is an example of a noun from the list being used for this purpose from the BNC Humanities corpus (text APS):

> '"I mean, there's no beginning, no middle, no end. There's no coherence" (Wright 1985, p. 266). As usual there are metafictional implications in ***this statement.'***

This additional example is from the same corpus (text EEY):

> 'Neither Elizabeth nor James, [Lloyd] said, had allowed the Duchy of Lancaster to be absorbed into the Exchequer… In making **this claim** Lloyd described very well one of the two important functions of early modern bureaucracy.'

In order to identify which items from the NICNGL are suitable for this purpose, I applied a qualitative criterion to the list, asking if the semantic properties of the word allowed it to act as a lexical reference to a citation. Further research targeted specifically to this area could establish other lexical items which identify additional lexis used for this purpose. The following nouns might be explicitly focused on during EAP input to illustrate their value as a resource for this purpose:

Table 6: *Items from the NICNGL useful for following citations*

Statement	Claim	belief	account	Result	evidence	point	problem
Information	concept	notion	fact	Distinction	argument	analysis	perspective

Items from the NICNGL available as resources for nominalizing processes

As stated in the literature review, once a concrete action in the real world has been described and is then referred to in a subsequent sentence, a nominalisation is often used. These nominalisations are often combined with a reference word to explicitly link to the previous idea. Here is an example of a process being nominalized from the BNC Social Sciences corpus (text AMG):

> 'Individuals take turns in sitting vigilantly alert while others feed**, thereby functioning as watchdogs or guards**. There is a regular changeover between individuals in the performance of **this activity**.'

A number of items from the NICNGL appear to be powerful resources for this kind of nominalizing process. The table below has been constructed by searching the NICNGL for items that exhibit the semantic properties required to abstract even complex actions and processes into a single term. EAP input could focus on the significance of these items for achieving cohesion and lexical density in texts, where a nominalisation can replace an entire process. Here are the key terms from the main list:

Table 7: *Items from the NICNGL available as resources for nominalizing processes*

practice	System	activity	change	development	method	approach	process

Here is an example of the kind of question that could be used to increase skill in using nouns as nominalized, lexical references to processes:

Since the introduction of a more democratic decision-making mechanism across the organisation, employees have reported much higher feelings of loyalty towards the company. This positive _____ in attitude may also lead to increased productivity. Which of these nouns best completes the second sentence? A) process B) method C) development

Semi-fixed, partitive lexical phrases

Given that the major purpose of producing the NICNGL is to identify the items that most frequently contribute to lexical cohesion and lexical density within academic texts, I have distinguished between nouns used as content words in cohesive nominal groups and those forming a partitive construction such as 'this form of' and 'this type of'. Here is an example of 'this kind of' used in this way from the BNC (text G0R):

> 'Why, in higher education, do we tend to associate **this kind of** intellectual **activity** more with the … Humanities … than with the pure sciences and the technologies?'

The table below shows the most frequent nouns in the combined corpora used to create partitive constructions. The second column shows the total number of times these nouns appear in the eight strings under investigation and the third shows the number of occurrences of the noun in [this + noun + of] structures in the combined corpora.

Table 8: *Nouns mostly used in cohesive, partitive lexical phrases*

Noun	Totals occurrences in the combined corpora	Total occurrences of [this + **noun** + of] structures with *kind*, *type*, *sort* and *form*
kind	489	450
type	266	194
sort	214	194
form	143	59

Whilst not content words, these structures are potential resources for EAP students in the construction of more complex nominal groups. In particular, they could be presented as semi-fixed lexical phrases (Lewis 1997:15), functioning as devices in a text for broadening and narrowing an analytic focus. The generalizing function of these semi-fixed phrases is noticeable in the following example in which 'this kind of activity' refers to a range of different kinds of activities:

> "… they also experience limited **opportunities for engaging in social routines** … [I]t may be because particular educational placements have very limited opportunities and few resources for **this kind of activity**. (BNC - CG6)

Since the first instance of the activities mentioned by the writer is very general, the subsequent reference to it is also rendered general with the structure 'this kind of [noun]'. As well as for generalising, though, these phrases are also used to narrow a focus. In these cases, 'this kind of' means 'this particular kind but not others'. In this way, these phrases function in a similar way to [such + noun] anaphoric, lexical references. The nouns following 'of' in these semi-flexible phrases are frequently items from the NICNGL and so there is an opportunity, in terms of course planning, to link the NICNGL and these cohesive partitive structures.

Other lexical phrases

Other nouns appearing with a high frequency in the eight strings in the combined corpora also form lexical phrases. As such, they appear to play more of a demarcated cohesive role in academic texts. Since they are not as frequently deployed as flexible lexical resources as the items on the NICNGL, they have not been included in the NICNGL. The nouns 'way', 'respect' and 'reason' appear frequently in the combined corpora as part of the lexical phrases 'in this way', 'in this respect' and 'for this reason'. These lexical phrases operate as cohesive discourse markers within texts and, although they are further towards 'fixed' in terms of their flexibility as lexical resources than the words in the NICNGL, they are valuable cohesive resources. Accordingly, these lexical phrases could be dealt with separately by EAP practitioners.

The following table shows the total number of instances of these nouns in the eight strings in the combined corpora along with the number of times they appeared as part of the lexical phrase in question:

Table 9: *Cohesive lexical phrases not included in NICNGL*

Total occurrences for '**way**'	Total occurrences for '**in this way'**	Total occurrences for '**respect**'	Total occurrences for '**in this respect'**	Total occurrences for '**reason**'	Occurrences for '**for this reason'**
719	477	228	214	180	124

The occurrence of these four lexical phrases within the data points to the fact that lexical phrases play an important role in academic discourse. It is interesting that in setting out to find [this + noun] collocations, among others, in academic discourse, lexical phrases also appear in the data. This is perhaps a confirmation of the approach to academic language development that prioritises collocation and lexical bundles over discrete vocabulary items, as suggested by Durant (2009).

Frequency of the eight structures investigated in the combined corpora

Across the combined corpora, there were a total of 36,136 occurrences of these eight forms of nominal groups. This below table provides EAP practitioners with data on the frequency with which these structures appear in professional academic writing. It suggests, for example, that when introducing sentence transitions by means of cohesive nominal groups, it may be salient to begin with 'this + noun' structures since they are likely to be the most familiar to students and relevant for their writing. [This + noun] structures can then be modified with adjectives and plurals in order to increase the range of resources available to EAP students. Given the significantly larger size of the Humanities corpus, this data also shows that eight the types of cohesive nominal groups investigated here appear more frequently in the Social Science corpus than the Humanities.

Although, as previously stated, Tagant was not able to distinguish between uses of 'that' successfully, I conducted a manual count of [that + noun] strings with a cohesive function. For this count, I used the concordance tool on lextutor.ca, using the 6m token general academic corpus, in order to provide some data in this area. For one hundred words drawn from the AWL, there were 149 instances of cohesive nominal groups formed with [that + noun]. For these same 100 academic words, there were 1742 instances of the [this + noun] clusters. This large difference indicates that [this + noun] clusters appear in the copora investigated at a ratio of almost 12:1 compared with [that + noun] clusters. Interestingly, the ratio between [these + noun] and [those + noun] is much closer. There were 609 instances of [these + noun] clusters and 139 cases of [those + noun] clusters evident in the same 100 AWL words investigated on lextutor.ca, resulting in a ratio of approximately 4:1. As a guide, these two additional structures have been included in the table below and are indicated in grey.

Table 10: *The frequency of the eight structures in the combined corpora*

Syntactic form of nominal group	Total number of occurrences in combined corpora	Soc Sci occurrenc-es	Human occurrences	Percentage of the total occurrences in combined corpora
This + noun	17999	8681	9318	50%
These + noun	7237	4037	3200	20%
Such + noun	4642	2108	2534	13%
This + adjective + noun	2556	1166	1390	7%
Those + noun [data from lextutor.ca]				
Such + a/an + noun	1667	708	959	5%
That + noun [data from lextutor.ca]				
These + adjective + noun	1098	637	461	3%
Such + adjective + noun	663	263	400	2%
Such + a/an + adjective + noun	274	117	157	1%
Totals	36,136	17,717	18,419	

The cost of a broad-spectrum word list to specific disciplines

There are a number of lexical items that did not meet the criteria for inclusion on the NICNGL due to their having a high frequency and/or range in only one of the corpora but not

in the other. The following table shows the significant items from each corpus that met the quantitative criteria in only one corpus:

Table 11: *Items meeting criteria in only one corpus*

Lexical items meeting criteria in Social Sciences corpus	research, model, analysis, data, finding, knowledge, result, task, trend, society, procedure, child, discussion, volume, strategy, technique, subject, topic, role, behaviour
Lexical items meeting criteria in Humanities corpus	year, people, passage, connection, text, poem

The items in Social Sciences section of the above table, in particular, seem highly desirable as lexical resources for students entering this field. Whilst EAP may be taught as a broad-spectrum discipline, it should not be forgotten that more situated language input may be required to prevent gaps in productive abilities appearing. Inevitably, a broad-spectrum word list such as the NICNGL excludes terms whose frequency and coverage grows in significance as the focus of the lens narrows. If more finely-tuned provision is available, I would recommend the inclusion of the Social Sciences sub-list in particular. However, as long as institutional factors tend to require all non-hard-science students to be taught together, general purpose lists such as the AWL and NICNGL will have an important role.

Further applications of the NICNGL for classroom practice

The use of corpora in class can provide teachers and students with insight into how language works in academic texts. As mentioned, the corpora on which this study is based are available for further investigation as part of the English concordance feature of lextutor.ca. Such investigations in class with lextutor.ca and other corpora may be significant in embedding and extending knowledge of cohesive practices within academic English into a student's lexicon. With this in mind, the NICNGL could be further exploited in class in the following ways:

1. The following methodology could be used to build knowledge of collocations within cohesive nominal groups. Students select a subset of nouns from the NICNGL which they wish to investigate further. They search for a noun in either the Humanities, Social Sciences corpus or General Academic corpus within lextutor.ca, sorting the results 2 words to the left of the search term. Students can then scroll through the concordance lines looking at the

adjectives occurring between the determiner and noun. For example, if you investigate [this + adj + process] in the Social Sciences corpus, the following adjectives are noticeable: natural, selective, first, last, second, last, among others. Later activities could focus meaningful practice of the discovered language.

2. Students can generate more subject specific lists of nouns used anaphorically using the Corpus of Contemporary American English (COCA) (Davies, 2008) and can compare them with NICNGL using the following method. They need to register for access to the corpus at http://corpus.byu.edu/coca/. Students then select a sub-corpus within the academic section of the corpus, e.g. medicine, using the search interface. Next, they investigate this body of language for nouns following the determiners 'this' and 'these' by adding the correct part of speech tag, e.g. 'this [nn*]' and 'these [nn*]'. They then scan the results for items pertinent to this particular subject, but not present in the NICNGL. This particular enquiry in the sub-corpus of medicine yields the following nouns among others: treatment, sample, procedure and technique.

3. The following methodology should raise awareness of the variety of syntactic structures used anaphorically in nominal groups in academic writing. A piece of text is provided to students with the 8 syntactic structures used in this study embedded, e.g. [this + adj + noun], [these + noun] and [such a + noun], etc. The students are asked to read the text identifying each instance of a reference back to something earlier in the text. In this way, each of the 8 syntactic structures is identified. Students discuss similarities and differences between these structures. Then, they speculatively place the 8 structures in order of how frequently they appear in academic text. This order is in checked using the COCA corpus by selecting the academic section of the corpus in the search interface and investigating it with the relevant search strings, e.g. [this + noun]', [these + noun] and [this + adj + noun], etc. Each time a search is conducted, the total number of occurrences in the corpus is noted in each case. For example, [this + noun]' yields 14111 results and [this + adj + noun] yields only 357. Finally, the students can check the actual frequency of these structures in COCA against their own speculative order.

The above activities, being somewhat technical, may not suit the learning preferences of all learners and, therefore, would probably benefit from being embedded in lessons allowing for additional communicative and productive output.

Conclusions

This study set out to provide a data-informed list of nouns frequently occurring in cohesive nominal groups in academic writing. Investigating the combined corpora of texts from the Humanities and Social Sciences has generated a list of around seventy lexical items (the NICNGL) that could be part of EAP input intended to facilitate the development of lexically dense and lexically cohesive L2 academic writing. Two significant sub-lists have been identified: a list of items particularly useful for comments following a citation and a group of nouns that nominalise processes. Four frequently occurring nouns in the data have been identified as forming partitive constructions with a cohesive aspect enabling the writer to narrow or broaden the range of analysis in the writing. In addition, there is a proposed order in which these cohesive structures could be introduced within a syllabus. However, in examining the data for lexical items that met the quantitative criteria in one corpus but not the other, it is apparent that the CICNGL, as a broad-spectrum word list, will not cover all of the resources required for localised academic disciplines. A number of key terms within the Social Sciences field, in particular, are recommended as a sub-list where more localised input is possible. Notwithstanding these findings, it would be valuable to conduct a similar kind of frequency and range analysis in other fields of academic study, such as the hard sciences. Additional research could also establish whether a similar study of American and other academic Englishes produced the same list of significant items. Further investigation of word frequencies in cohesive nominal groups might also establish the most frequent [adjective + noun] collocations.

Biodata

Andrew Drummond teaches English for Academic Purposes at King's College London. His research interests include cohesion in academic writing, corpus linguistics and academic lexical development.

References

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Anthony, L. (2015). TagAnt (Version 1.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. University of Sydney Papers in TESOL, 5(5), 31-64.

Cobb, T. (2015) Corpus Concordance English. Available from: http://www.lextutor.ca/conc/eng/. Accessed 10/12/15

Coxhead, A. (1998). An Academic Word List. ELI Occasional Publications #18, School of Linguistics and Applied Language Studies, Victoria University of Wellington: Wellington.

Coxhead, A. (2000). A new academic word list. *TESOL quarterly*, 213-238.

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, *33*, 66-76.

Davies, M. (2008). The Corpus of Contemporary American English: 520 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.

Drummond, A. (2015). How IsiZulu speakers use cohesion in English in their academic writing. Unpublished Master's Thesis. School of Education, Faculty of Humanities, University of Witwatersrand, South Africa.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. English for Specific Purposes, 28(3), 157-169.

Fang, Z., Schleppegrell, M. J., & Cox, B. E. (2006). Understanding the language demands of schooling: Nouns in academic registers. *Journal of Literacy Research*, *38*(3), 247-273.

Halliday, M., & Hasan, R. (1976). *Cohesion in English*. Routledge.

Halliday, M. (1994). Around the clause cohesion and discourse. (pp. 287-318). *An introduction to functional grammar*.

Halliday, M., Matthiessen, C. M., & Matthiessen, C. (2004). *An introduction to functional grammar*. Routledge.

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"?. *TESOL quarterly*, *41*(2), 235-253.

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: a developmental perspective. *Working Papers in Linguistics*, *53*, 61-79.

Kaplan, R. B. (1967). Contrastive rhetoric and the teaching of composition. *TESOL quarterly*, *1*(4), 10-16.

Lautamatti, L. (1987). Observations on the development of the topic in simplified discourse. In U. Connor & R. B. Kaplan (Eds.), *Writing Across Languages: Analysis of L2 Texts* (pp. 87-113). Reading, MA: Addison-Wesley

Lewis, M. (1997). *Implementing the Lexical Approach: Putting Theory in Practice*. Hove: Language Teaching Publications

Mueller, B. (2015). *Analysis of Nominalization in Elementary and Middle School*

Science Textbooks. Paper 247. MA Thesis. Hamlime University. Retrieved from: http://digitalcommons.hamline.edu/cgi/viewcontent.cgi?article=1246&context=hse_all

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In *Vocabulary in a second language: Selection, acquisition, and testing*, 3-13.

Nation, I. S. P. (2016). *Making and Using Word Lists for Language Learning and Testing.* John Benjamins Publishing Company.

Google Books Ngram Viewer. (2016) Google. Search terms: 'furthermore, notwithstanding. Available at: https://books.google.com/ngrams

O'Keeffe, A. (2000). Undergraduate Academic Writing: an analysis of errors and weaknesses in syntax, lexis, style and structure. *In Language and Literacy for the New Millenium* (pp. 167-186). Dublin: Reading association of Ireland,

Sadighi, F. (2012). Cohesion analysis of L2 writing: The case of Iranian undergraduate EFL learners. *Mediterranean Journal of Social Sciences*,*3*(2), 557-573.

Schmitt, D & Schmitt, N. (2005). Focus on Vocabulary: Mastering the Academic Word List. Longman

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

West, M., & West, M. P. (Eds.). (1953). *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. Addison-Wesley Longman Limited.