Vocabulary Learning and Instruction
Volume 6, Issue 1, November 2017
http://vli-journal.org

Responding to Research Challenges Related to Studying L2 Collocational Use in Professional Academic Discourse

Birgit Henriksen and Pete Westbrook
University of Copenhagen

doi: http://dx.doi.org/10.7820/vli.v06.1.Henriksen

Abstract

This study describes the English collocational use of non-native university teachers from two different disciplines lecturing in an English-medium instruction context at the University of Copenhagen (UCPH). The primary focus is on how we addressed the research challenges involved in identifying and classifying collocations used by L2 speakers in advanced, domain-specific oral academic discourse. The main findings seem to suggest that to map an informant's complete collocational use and to get an understanding of disciplinary differences, we need to not only take account of general, academic and domain-specific collocations but also need to cover the full range of both lexical and grammatical collocations.

1. Introduction and Background to the Study

1.1 The Collocational Use of Non-Native Lecturers Teaching in an EMI Context

The past 20 years or so has witnessed rapid growth in internationalisation at institutions of higher education in Europe, especially northern Europe, not least in Denmark. As a result, English has increasingly become the lingua franca of academia, as more and more degree programmes are run through the medium of English. Consequently, university teachers who are non-native speakers of English are asked to lecture, tutor and supervise students in English and are thus expected to perform effectively in professional academic discourse in their L2.

To meet the pedagogical challenges presented by this situation, the University of Copenhagen implemented a language policy based on the parallel language use of Danish and English. To support this internal language policy, as well as to ensure quality in educational programmes and research, the University established the Centre for Internationalisation and Parallel Language Use (CIP) in 2008 as a research, competence development and resource centre. Part of CIP's remit was to provide language training and language certification of tenured academic staff. As a result, the performance-based Test of Oral English Proficiency for Academic Staff (TOEPAS) certification procedure was developed by the Centre and is used for assessing whether university lecturers have sufficient oral proficiency for

Please cite this article as: Henriksen B. and Westbrook P. (2017). Responding to Research Challenges Related to Studying L2 Collocational Use in Professional Academic Discourse. *Vocabulary Learning and Instruction, 6* (1), 32–47. doi: 10.7820/vli.v06.1.Henriksen.Westbrook

coping with the communicative demands of English-medium instruction (EMI) (see http://cip.ku.dk/english/certification/).

This paper reports on a study of university teachers' L2 collocational use when lecturing in an EMI context and is based on data from the TOEPAS certification. Fifteen of the mini-lectures from three different academic domains were transcribed for research purposes. The authors set out on a small exploratory project which aimed to describe the lecturers' *overall collocational use* across all the collocational sub-types that would be expected to be found in academic language, that is, domain-specific, academic and general collocations. The aim of this original study was to test whether there were any parallels between the lecturers' level of English proficiency as assessed in the certifications compared to the frequency and appropriateness of collocational use across the three types of collocations mentioned above. Moreover, we wanted to identify possible similarities and differences in collocational use across different academic domains. By including all three types of collocations in our analysis, we would be able to generate a general score of informants' collocational use which could be used to correlate with other measures like fluency, certification score and vocabulary size measures.

The aim of the present paper is to highlight the research challenges inherent in investigating collocational use in oral, domain-specific language, both in general, but more specifically across the three sub-types mentioned above; an area which seems to have been subject to very little research as far as we are aware. We will present suggestions for dealing with these challenges and the effect of the choices made on the results. Although this is a small-scale study, we believe it can contribute to knowledge about collocational use in academic discourse, particularly on how this could and should be researched.

1.2 The Importance of Domain-Specific, Academic and General Collocations in Academic Discourse

Collocations are frequently recurring two-to-three-word syntagmatic units (e.g. *soft noise, tolerance for*). In the research literature, collocations are defined as a subset of formulaic sequences, distinct from other types of formulaic sequences such as lexical bundles, idioms, and pragmatic phrases (Nattinger & DeCarrico, 1992). Handl (2009) sums up collocations thus: "We can conclude that collocations are *conventionalized* recurring word combinations exhibiting more or less *restrictedness*, more or less *semantic opacity* and a certain degree of *predictability* for native speakers (…). So, two words that collocate are not governed by semantic compatibility, but rather by lexical restriction, that is, by the norms of the language" (Handl, 2009, p. 70, our italics). As examples of this, we can take the combinations *strong coffee* and *powerful car* as the preferred collocates, rather than *powerful coffee* and *strong car.*

Collocations can consist of different grammatical and lexical constituents. A lexical collocation is the type of construction where a verb, noun, adjective or adverb forms a predictable connection with another word from these word classes, as in *completely satisfied* (adverb+adjective), *excruciating pain* (adjective+noun) and *commit suicide* (verb+noun). A grammatical collocation is a type of construction where for example a verb or adjective must be followed by a particular preposition, as in *depend on* (verb+preposition) or *afraid of* (adjective+preposition).

Mastery of formulaic sequences, including collocations, has been described as a central aspect of communicative competence, enabling the native speaker to process language both fluently and idiomatically and to fulfil basic communicative needs. Nation (2001, p. 318) concludes that "all fluent and appropriate language use requires collocational knowledge." It has also been argued that collocational use is equally important for L2 learners (Barfield & Gyllstad, 2009; Henriksen, 2013). Nevertheless, this is a language phenomenon which is said to be acquired late and which is often not mastered very well even by reasonably competent L2 language learners (Nesselhauf, 2005; Laufer & Waldman, 2011; Henriksen, 2013). For this reason, collocational proficiency may be seen as a quality feature of advanced language use, for example, for academic lecturers like our informants who are operating in a highly demanding professional setting in their L2.

The main reason for focusing on collocations in relation to EMI language use is that collocations typically have a highly referential function (Howarth, 1998), as opposed to the discourse or pragmatic functions of other types of formulaic sequences. Moreover, they tend to be very genre specific. As such, collocations are often seen as characterising technical sub-languages (Ananiadou & McNaught, 1995), that is, languages from different study domains. Similarly, mastery of collocations may be a hallmark of certain types of academic writing which emphasize clarity, precision and lack of ambiguity (Howarth, 1998). As mentioned, even very advanced L2 users seem to have problems with using collocations and, apart from leading to unfortunate misunderstandings, advanced non-native speakers' collocational deviations may signal a lack of academic expertise (Henriksen, 2013, p. 37).

In the research literature, vocabulary is divided into general, academic and technical language (Coxhead, 2000; Hwang & Nation, 1995; Xue & Nation, 1984). However, this research focuses very much on single word items, with little or no research on the same distinctions applied to collocations. Hwang and Nation (1995) found that vocabulary in non-fiction texts can be divided into high frequency (or general service) vocabulary, sub-technical (or academic) vocabulary, technical (or domain-specific) vocabulary, and low-frequency vocabulary (based on Nation, 1990, p. 19). How these different categories of items combine often characterises different kinds of professional academic discourse. Because of the complexity of professional academic discourse found in the certification data, the authors also found it necessary to make the distinction between general, academic and domain-specific collocations in line with the single word item distinctions brought out by Hwang and Nation (1995). A study by Westbrook (2015), who investigated the role of collocations for fluency in the same data set as used in the present paper, found significant differences in results depending on whether domain-specific collocations were included in the calculations or not. This seems to be in line with differences in the density of domain-specific single words between disciplines found by Chung and Nation (2004), and would therefore present a case for also distinguishing between general, academic and domain-specific collocations in our study of academic discourse.

1.3 The Research Issues Addressed

As pointed out by Henriksen (2013), virtually all the previous studies on collocations have dealt with general collocations with only very few, if any, tackling academic and technical, that is, domain-specific collocations. In addition, most research on collocations has focused on written or corpus data. Moreover, these studies have often been limited to one specific type of collocation, typically verb+noun or adjective+noun collocations. By including a range of collocational sub-types used in oral, academic discourse, we have been expanding the range of research focus. The lack of studies dealing with various types of collocations, however, meant that there were very few previous comprehensive research models to draw on in our analysis of our informants' overall collocational use.

The extensive pilot phases, which have been reported at various conferences (Complexity and Idiomaticity, Stockholm University, June 2012; EIE Conference, Copenhagen 2013; SDU SELC Conference, Odense 2013; PhD Applied Linguistics (Lexical Studies) annual conference, Cardiff University, Wales 2014; and AILA World Congress 2014), highlighted a range of methodological problems, both in relation to deciding how to operationalise the distinction between domain-specific, academic and general collocations and how to identify these three types of collocations in the data. Moreover, our preliminary studies showed that the internal structure and complexity of the individual collocations seemed to differ across the three types of collocation, creating analytical challenges in relation to how to deal with more complex, embedded collocations and what to include in the quantification of the individual collocations. In addition, the analysis of oral data created its own challenges, for example, in relation to split collocations, where the distance (span) between node and collocate (Nattinger & DeCarrico, 1992) was in some cases quite considerable (see Section 3.2).

All these research challenges, which needed to be overcome in order to carry out robust research on collocations in domain-specific academic discourse, prompted us to write the current paper. This paper is therefore concerned with identifying the methodological problems and investigating how results might differ according to the methodological choices made. The study includes an analysis of 12 CIP TOEPAS lectures from two academic domains. The first two research questions are related to the research procedure itself:

(1) What challenges are there in trying to describe collocational use across general, academic and domain-specific types?
(2) How might these challenges be met?

On the basis of our suggestions for solving these research challenges, we will present the results of the analysis of the 12 lecturers' collocational use to answer the last three research questions. The focus here is on exploring potential differences across the different collocational types and across the two academic domains:

(3) What characterises academic collocational use across lexical and grammatical collocations?

(4) What characterises academic collocational use across the general, academic and domain-specific categories of collocations?
(5) What characterises collocational use across different academic domains?

2. Previous Research on Collocations

The initial research challenge for any study on collocations is to decide on a method for identifying and delineating the types of collocations to be explored. This question will initially be discussed in a short literature review on collocational research in this section. The final research choices made will be outlined in the actual presentation of the study itself in Section 3.

Two main approaches have been adopted by researchers to identify collocations in a given text or corpus: the *frequency-based approach* and the *phraseological approach*. The frequency-based approach is associated with computer-based searches of large language corpora. These searches involve identifying words that occur within a short span, usually four words, either side of a headword, or "node." If the node occurs together with another word or words within this span "at a frequency greater than chance would predict, then the result is a collocation" (Nattinger & DeCarrico, 1992, p. 20). Thus, collocations are not necessarily contiguous, although they can be. They can also be realised in different lexical combinations. The collocation *strong argument*, for example, can be realised as: *it is a strong argument, he argued strongly for, the argument is a strong one*, and so on. Frequency criteria alone, however, will not necessarily yield all possible collocations. The phraseological approach employs a manual identification based on more intuitive syntactic and semantic analysis of word combinations and is helpful in defining collocations more precisely. Generally, researchers have adopted a combined approach (Barfield & Gyllstad, 2009).

The next fundamental challenge in any study of collocations is deciding what types of word combinations to include as collocations. One question is whether or not to include compounds. Granger and Paquot (2008) argue for excluding them in any analysis of collocations, partly because of their "uncertain status as single or multi-word units" (Granger & Paquot, 2008) (e.g. *good will, good-will, goodwill*), and partly because of their somewhat fixed status (e.g. *black hole, goldfish, blow-dry*). However, domain-specific discourse tends to be compound noun heavy. As Moon (1997) states: "compounds typically denote and have high information content – often because they are technical terms or have specific reference" (Moon, 1997, p. 56). Such compounds tend to be more flexible than Granger and Paquot claim; there is, after all, to take an example from our data, a difference between a "mouth speculum" and an "ear speculum." In addition, they are also included in such collocational reference works as the Oxford Collocation Dictionary (OCD) and in Pearson's Academic Collocation List (ACL). In a study of collocational use including general, academic and domain-specific collocations, the inclusion of compounds would therefore seem to ensure that the full range of collocational types would be represented.

A third point to consider is related to the distinction between lexical and grammatical collocations outlined in Section 1.2. Most research studies on collocations have investigated a limited range of collocational types, often with a focus on lexical collocations (Henriksen, 2013). The broad categories of lexical and grammatical collocations can be further broken down into different structural sub-types in relation to the different word class constituents they include. As mentioned above, most research on collocations seems to have focused only on the adjective+noun and the verb+noun constructions (Henriksen, 2013). If the aim is to describe informants' overall collocational use, both lexical and grammatical collocations types and the different sub-types would, however, need to be included in the analysis. As will be documented later, academic language includes a range of structural combinations of both lexical and grammatical collocations, for example, adverb+verb (*gradually realize*), noun+preposition (*idea of*), verb+adjective (*be surprised*), verb+noun (*take time*) and verb+preposition (*ask about*).

Apart from the basic structural collocational types mentioned above, there are also other more complex collocational combinations, "nested collocations" (Frantzi & Ananiadou, 1996), which consist of a collocation in combination with additional word class constituents. These complex collocations, which can include both lexical and grammatical sub-types, are typically found in domain-specific discourse but have not been the focus in collocation research in general so far. The types found in our study include the nested collocational combinations shown in Table 1.

The final challenge is related to the distinction between general service (*spend time, good idea, hear about*), academic (*strong argument, reliable data*) and domain-specific (*exponential bounds, be contained*) collocations, which will be one of the main topics for the rest of this paper. Due to the existence of collocations, dictionaries such as the OCD, and with the recent publication of Pearson's Academic Collocations List (Ackermann & Chen, 2013; http://pearsonpte.com/research/academic-collocation-list/), general and academic collocations are relatively simple to identify and delineate in an objective way, whereas identifying and classifying the domain-specific collocations is far more challenging. A few studies have investigated mathematical and medical collocations (e.g. Haag, Heppt, Stanat, Kuhl, & Pant, 2013; Herbel-Eisenmann, 2002; Méndez Cendón, 2004), but it is very difficult to find an objective method for identifying and classifying domain-specific collocations which can be used across different academic domains. Often domain-specific or technical language is associated with the use of specific "technical terms" or "phrases," for example, highlighted in domain-specific dictionaries or terms lists, but no standard method for establishing these inventories have been developed, and many academic fields have not developed or published lists of domain-specific collocations.

Table 1. Lexical and Grammatical "Nested Collocations"

	Collocational combination	Examples from the data
Lexical nested collocations	adjective + collocation	*complicated differential equation*
	adverb + collocation	*purely algebraic definition*
	collocation + collocation	*finitely generated abelian groups*
	collocation + noun	*solutions to this equation*
	proper noun + collocation	*Heisenberg's matrix mechanics*
	verb + collocation	*have a continuous function*
	adjective + collocation +collocation	*compact Hausdorff topological space*
	noun + collocation	*capilliary refill time*
Grammatical nested collocations	collocation + preposition	*continuous functions on*
	preposition + collocation	*in a physical world*

3. Methodology

3.1 Data Source

The TOEPAS test is a high-stakes oral assessment and takes the form of a 20-minute simulated mini-lecture in English, carried out at the University of Copenhagen (http://cip.ku.dk/english/certification/). Teachers are assessed on a 5-point holistic scale based on five dimensions (pronunciation, grammar, lexis, fluency and interaction skills). Scores 3, 4 and 5 are certified, while 1 and 2 are not certified. Teachers come from different faculties and, as part of the test procedure, are given formative feedback as well as their overall score. Each mini-lecture is videoed and CIP now has a databank of around 400 certifications. The data for this paper have been collected from 12 lecturers from two different departments: the Department of Large Animal Science (LAS) and the Department of Mathematics (Maths). The scores for the six LAS informants were 5, 4, 3, 3, 2, 2, and for the six Maths informants 5, 4, 4, 4, 3, 2, respectively.

3.2 Identifying Collocations in Our Data Electronically or Manually?

One possible method for identifying potential general, academic and domain-specific collocations in our data set using the frequency approach is to apply corpus-based tools such as WordSmith or AntConc. Among other things, these tools allow the user to list words in the text in order of frequency, show frequent word partnerships present in the texts, and display concordance lines for a particular word, with the additional option of sorting the lines according to the co-text to the left or right of the node word. Thus, the programmes enable the identification of clusters of word partnerships which form potential (domain-specific) collocations. In addition, if the data set is large enough, both Wordsmith and AntConc have the function to compare the data texts with another (general) corpus in order to identify which word occurs more frequently in the data than in a general data set (the so-called “keyword analysis”). The advantages of using electronic corpus-based tools are that they are objective, apply clear criteria and are less time-consuming. However, utilizing such electronic tools was not a viable option for us, as our corpus of 12 mini-lectures was simply too small and we had no access to a reference corpus for the separate domains. Moreover, an electronic search would not have been able to cope with the long span in “split” or “fragmented” collocations (Pulverness, 2007). These are collocations where the span between the node and collocate is quite long, for example, *the path that a ball rolling down the hill would take*. The varying lengths of the domain-specific collocations (typically from two- to five-word units) would also have made it difficult to work electronically. Finally, we would still need to manually identify and discard the non-meaningful lexical bundles, for example, *and the*, *as we,* etc. (Martinez & Schmitt, 2012) from the potential collocations.

Despite being more subjective and time-consuming, a manual search therefore had the advantage of allowing us to catch all the potential collocations in the data transcriptions, and take account of the problem of “split collocations” as in the example mentioned above. Many of these constructions were found in the oral data, due to the specific features of the oral discourse mode. The manual approach also ensured that we would identify and count two (or more) collocational pairings on the same headword. These were counted as two separate collocations. For example, *use a different method* was counted as two collocations (*use a method* and *different method*).

Thus, our final approach was based on the manual/phraseological approach. This involved each of the authors identifying all potential collocations manually simply by going through each of the 12 transcripts independently and underlining all the potential collocational units. To ensure that as many potential collocations as possible could be underlined, especially the domain-specific ones that we had no expert knowledge of, any unknown combination which in any way could be construed as a collocational unit was included in the initial identification procedure. We then compared those we had found individually. In most cases, we found the same, but there were also a certain number found by one researcher and not the other, which demonstrates the advantage of having two coders working on the same data, especially in order to "empty" the data in an attempt to identify as many potential collocations as possible. This initial identification stage is described as stage one in Figure 1.

3.3 Categorising and Delineating General, Academic and Domain-Specific Collocations in Our Data

Based on extensive pilot coding, we finally settled on a combined method, adopting a three-tier approach of exclusion: first classifying the academic collocations using Pearson's ACL (stage 2), then classifying the general collocations using the OCD (stage 3) and finally identifying the domain-specific collocations by checking them through a Google search (stage 4). These gradual

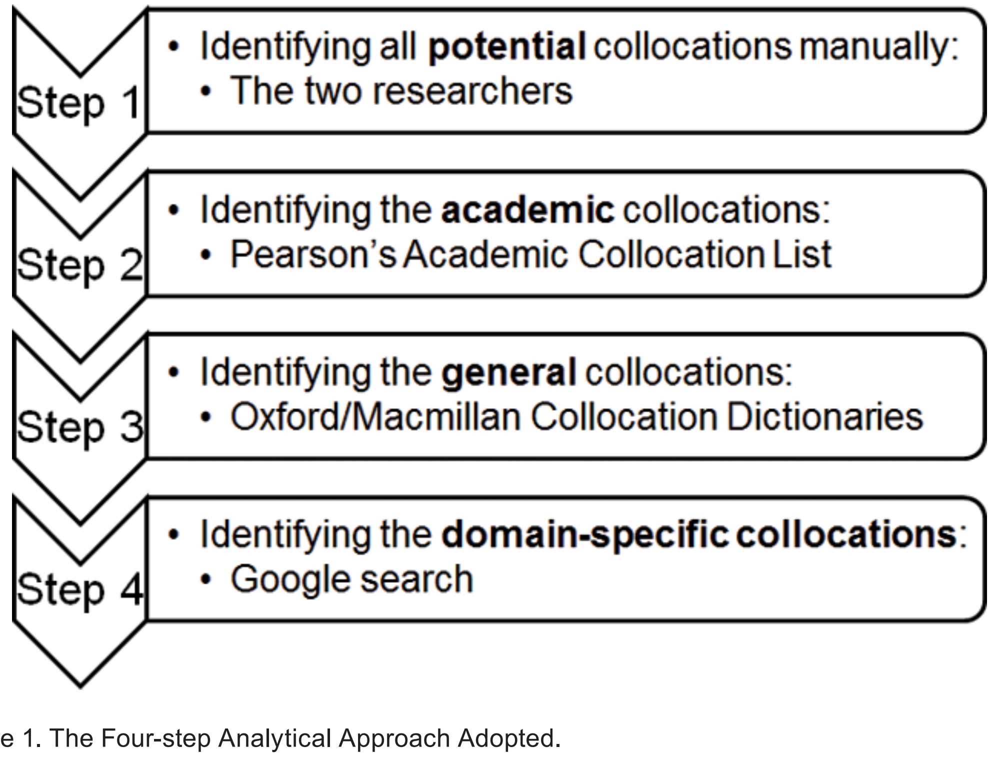
Step 1
• Identifying all **potential** collocations manually:
• The two researchers
Step 2
• Identifying the **academic** collocations:
• Pearson's Academic Collocation List
Step 3
• Identifying the **general** collocations:
• Oxford/Macmillan Collocation Dictionaries
Step 4
• Identifying the **domain-specific collocations**:
• Google search

Figure 1. The Four-step Analytical Approach Adopted.

stages of exclusion are illustrated in Figure 1. The more specific issues related to decision stage 4 for identifying the domain-specific types will be outlined in more detail later.

As mentioned above, lists of general collocations (e.g. the OCD) and lists of academic collocations (e.g. Pearson's list) were available tools for identifying and classifying the general and academic collocations. Unfortunately, the same type of research-based tool for identifying domain-specific collocations within our specific domains does not exist. One method of checking for domain-specific collocations could therefore be using expert raters (Chung & Nation, 2004) within a certain academic domain. In an initial pilot, three native speaker informants from LAS were asked to identify potential domain-specific collocations in the data from their own department. However, this procedure proved to be very time-consuming, both in relation to finding the informants, training them to understand the concept of collocations and having them deal with extensive lists of potential collocations. On top of this, the fact that it was also very difficult to obtain consensus from all three native speaker experts was another deciding factor for excluding the use of native speaker raters for the identification of domain-specific collocations. Thus, we decided to investigate other more objective methods of determining domain-specific collocations.

Initial pilot coding of the domain-specific collocations using different technical dictionaries and term lists showed that it would be very difficult to find a reliable procedure that could be applied across academic domains. The quality of the resources and the nature of the lexical entries created differences in the coding that would have seriously biased the results. Some resources primarily included single word items and noun+noun or adjective+noun collocations and very few listed collocations containing verbs or adverbs. The Google search procedure, that is, a frequency-based identification method, was therefore used to identify domain-specific collocations. This was done by putting the potential collocation itself in quote marks and then combining this search with a consistent phrase relating to each of the departments in turn. After several trial runs, it was decided to use the word "maths" for the mathematics informants and "animal science" for the informants from the LAS department, and the cut-off point was set at 5,000 hits. From the resulting combinations found, we checked for lemmas and added those in as necessary; in addition, we excluded any combinations which were clearly not domain-specific collocations (e.g. *background story*).

Finally, any collocations included in stage 1 as potential collocations that were not coded as academic, general or domain-specific were not counted as collocations at all and were excluded from the inventory of collocations found in our data set.

Hwang and Nation (1995), focusing on individual items, have stressed that we cannot operate with clear-cut distinctions between the different vocabulary types, but are dealing with a continuum with fuzzy and arbitrary divisions. This, however, raises the question of how and where the arbitrary dividing line can be most sensibly drawn and how to decide the classification of the individual collocations. Working with this three-tier procedure of gradual exclusion, we still encountered some problems related to the fuzzy boundaries between the categories. For the collocations coded as general, some of these may be pre-technical (e.g. *complex numbers, relations between*), while others may be crypto-technical, that is, polysemous word with one meaning clearly related to specific fields of study and where the technical meaning is not likely to be known by a lay person (Fraser, 2006), for example, *product of*, *be unique*. This is a well-known issue for "the language of mathematics," that is, the fact that maths words can have multiple meanings, for example, *true* in the general sense and a second more technical meaning of the word. The setting of the boundaries between the categories is also an issue when dealing with the collocations that were coded as domain specific. Some may belong to a larger group of study domains, for example, maths and animal science, the STEM areas: science, technology, engineering and maths (e.g. *do this procedure, subgroup of*), but they are not frequent enough across a range of academic domains to be listed as academic collocations, for example, in the Academic Collocations List.

3.4 Quantifying Our Results and the Issue of "Nested Collocations"

A further research issue to highlight is the question of what counts as one collocation. This is important as it is related to the quantification of the results. Technical language is often characterised by terms that are made up of multi -layered, nested collocations (described in Section 2), where an adjectival specification is added to a technical collocation, for example, *infinite dimensional space* or *symmetric unbounded operators*. Some are only used in the long version, whereas others are used by the same informant in multiple versions. We wanted to ensure that we were not analysing the nested collocations which were only found in the longer version as a combination of two separate collocations. Thus, our acid test was that if a combination only existed as a four-word combination in the informants' lecture data, it was counted as one collocation which was four words long. However, if components of the four-word combination were also used as, for example, two-word combinations, these were counted separately. For example, *infinite dimensional spaces* were found only in this form and were therefore counted as one three-word collocation. Conversely, both *unbounded operators* and *symmetric unbounded operators* were found in the data and were therefore counted as two separate collocations. To check which word combinations should be considered the "node" of the collocation, as a rule of thumb we counted the words in the collocation working back from the back (right) and working left (towards the front) (e.g. *value problem* with 614,000 hits and *initial value problem* which got 89,000 hits). This identified the combination "value problem" as the node for this complex nested collocation which was then coded as adjective+collocation and not as collocation+noun. All collocations were checked electronically (using AntConc) to confirm the number of instances of each collocation in the data.

4. Preliminary Results from Maths and LAS

Following the analytic procedures outlined above, we have so far finished the analysis of data from the 12 informants from LAS and from Maths.

As can be seen in Table 2, the two groups of informants produce more or less the same number of collocations (1776 and 1773). A similar picture emerges when we look at the collocational density per 1,000 words spoken (88 and 95, respectively). Differences, however, emerge when we look at the distribution between lexical and grammatical collocations, with 72% lexical and 28% grammatical collocations produced by the LAS informants, and 66% lexical and 34% grammatical collocations produced by the Maths informants.

Looking at Table 3, there were also a few other interesting findings related to some of the structural types used across the two groups. Considerably more noun+noun constructions were used by the LAS informants (15% compared to 2% for the Maths group), for example, *animal welfare, body mass, contrast effect* (LAS). More adjective+noun constructions were used by the Maths informants (42% compared to 31% for the LAS group), for example, *deep theorem, simple fact, algebraic operations* (Maths) and twice as many adverb+adjective constructions were used by the LAS informants as the Maths informants (10% compared to 5%), for example, *highly efficient, very bad, very noisy* (LAS). Regarding nested collocations, a considerably higher percentage were used by Maths informants (9%) compared to LAS (1%). Examples include *compact Hausdorff space, use the matrix norm, classification theorems for* (Maths).

The inclusion of all structural types made it possible to explore differences across the departments studied. For example, if noun+noun compounds had not been included, a specific feature of collocational usage differentiating LAS (193 instances) from Maths informants (40 instances) would have been missed.

Looking at the distinction between the general, academic and domain-specific collocations, other interesting differences between the two academic domains included in our study were found. As can be seen in Table 4, one in five (22%) of the LAS collocations were deemed to be domain specific, whereas over half of those (52.3%) identified in the Maths data were classified as domain-specific. If the domain-specific collocations had not been included, an important distributional difference in collocational use between lecturers from the two academic fields, LAS and Maths, would have been missed. Overall, both groups produced surprisingly few academic collocations. This is probably due to the use of Pearson's list which has been extracted from a written corpus, and may not reflect the usage of collocations in spoken academic communication. As shown by Dang (2016), an academic word list for oral language for single word items is different from Coxhead's list (2000) developed from a written corpus. Unfortunately, as far as we know, an Academic Collocations List for oral data is yet to be developed.

Table 2. Total Number of Lexical and Grammatical Collocations Used

	Large animal science group (N=6)		Mathematics group (N=6)	
	No. of collocations	% ages	No. of collocations	% ages
Lexical	1281	72	1179	66
Gram	495	28	594	34
Totals	1776	100	1773	100

Table 3. Breakdown in Lexical Collocations Used per Department

Structural types	Large animal science (*N*=6)		Mathematics group (*N*=6)	
	No. of collocations	% ages	No. of collocations	% ages
adj+adv	0	0	1	0
adj+n	396	31	493	42
adj+n+n	1	0	0	0
adj+v	1	0	1	0
adv+adj	132	10	59	5
adv+adv	2	0	0	0
adv+v	7	1	7	1
n+adj	2	0	0	0
n+n	194	15	40	2
n+phr	2	0	1	0
n+v	16	1.5	20	2
phr+n	12	1	5	0.5
pn+n	7	1	25	2
v+adj	184	14	175	15
v+adv	32	3	8	1
v+n	280	22	241	21
v+phr	1	0	0	0
v+v	1	0	0	0
Nested colls	11	1	103	9
Total lexical collocations	1281	100	1179	100

phr = phrasal verb.

Table 4. Collocations Used per 1000 Words Spoken Split into Academic, General and Domain-Specific Collocations

Collocational types	Large Animal Science (*N*=6)		Mathematics (*N*=6)	
	Per 1000 words	% ages	Per 1000 words	% ages
Academic	1.5	2	1	0.7
General	67	76	44	47
Domain-specific	19.5	22	50	52.3
Totals	88	100	95	100

5. Discussion and Perspectives

5.1 Research Questions

Two of the research aims of the project involve identifying the challenges related to studying collocational use in oral academic discourse and establishing a "one-size fits all" research methodology for researching collocations across academic disciplines. At present, the research outlined has revealed a number of research challenges that we have tried to meet by applying a manual procedure of identifying potential collocations and a three-tier analytical approach of mutual exclusion for classifying collocations as general, academic or domain-specific collocations. On the basis of this procedure of analysis, the data from our two study domains have been analysed and described.

As regards research questions 3, 4 and 5, the above results have revealed interesting differences in collocational use across the two domains investigated. The results highlight the importance of applying an extensive approach which enables us to achieve an overall measure of collocational use based on the inclusion of all collocational types found in the data. The grammatical collocations make up a sizeable proportion of the total, and their distribution varies across the academic domains. If, in line with many previous studies, we had only focused on for example the lexical collocations, we could have missed between a quarter and a third of the collocations used by our informants. Only including the general collocations, which have been the focus of most previous research on collocations, would have yielded a picture where the LAS informants are perceived to use more collocations overall than the Maths informants, which as we have seen is far from the case.

Looking more closely at the distinction between lexical and grammatical collocations, some slight differences between the relative proportions of lexical and grammatical collocations between the LAS and Maths informants were found. Including more informants that would have allowed for statistical analysis of the data may have revealed that these tendencies in fact describe statistically significant differences across domains. Interesting differences in the structural sub-types types used have also been found. For example, excluding noun+noun compounds would have omitted a large proportion of collocations (in the case of the LAS group, 11% of all collocations) and would have hidden another important difference between the profiles of the two groups (as only 2% of the Maths group's collocations were categorised as noun+noun compounds). As can be seen from the tables in Section 4, the distributional difference in the number of general and domain-specific collocations is also striking. Extending the study to include our three IT informants (which completes our data set) or even non-STEM disciplines, that is, the social sciences and humanities (SSH) may reveal further differences.

Looking at the general, academic and domain-specific collocations, it was found that the proportion of domain-specific collocations was considerably higher in the Maths group than in the LAS group, indicating that Maths may be a more "technical" discipline. Our results thus support the idea that clear differences across academic domains may be found in relation to the use of technical vocabulary. Chung and Nation (2003) found that an anatomy textbook contained a significantly larger proportion of fully technical (single word) terms (one in three) than an applied linguistics textbook (one in five). Fraser (2006) even found a figure of 35.9%, that is, more than one in three, technical words in a pharmacology text. In addition, many of the domain-specific collocations found in our data from the Maths group were made up of nested collocations. Both these phenomena seem to be because many of the terms used in Maths are made up of complex names describing particular theorems or modified names of theorems, for example, *Hausdorff space/compact Hausdorff space, C star algebra/concrete C star algebra/finite dimensional C star algebra*. In contrast, the domain-specific terms used by the LAS group tend to be more "penetrable" and stable, typically two-word noun+noun combinations such as *animal behaviour, bird predators, quail chickens,* and *pinyon jay*.

5.2 Future Research and Pedagogical Perspectives

Future research will include covering more disciplines, in order to make more nuanced analyses of potential differences in domain-specific academic language across disciplines. We will apply the same methods to analyse the data from the IT group of informants, and this will give us more comprehensive overall results related to possible differences across domains, perhaps revealing an even more complex picture of cross-disciplinary language use. Moreover, the inclusion of an additional discipline will reveal the robustness of our analytical approach. Finally, we will try to find a method of using expert raters to validate our initial categorisation of the domain-specific collocations which has been based on our Google search procedure.

In describing L2 learners' overall collocational use, Westbrook (2015) revealed that only focusing on general collocations was insufficient to show expected correlations between collocational use and fluency; however, such correlations may conceivably have been in evidence if the other types of collocations had been included. When results for the three domains are in place, we would like to run Westbrook's (2015) fluency measures against our data to find out if, with the inclusion of domain-specific collocations, any correlations can be found between collocational use and fluency. Including all the collocations used by the individual informant has made it possible for us to draw up various measures of collocational use that can be correlated with such fluency measures, but also other proficiency measures in forthcoming studies.

In the future, it would also be important to develop tools that may guide non-native students and researchers, especially for the non-maths domains, where fewer resources (such as technical dictionaries) may be found. This could involve developing lists of domain-specific collocations for various disciplines based on analysis of both L2 and L1 data. It could also be useful to develop sub-lists of clusters of collocations, for example, for frequent de-lexical verbs (have, get, give, make, etc.): HAVE + *a family of curves, the product of, a representation, a unique trace, a norm, a time evolution, real space*, etc. (see Menon & Mukundan, 2010). Finally, it could be useful to draw up a list of grey-zone area collocations, for example,. collocations that have both a general and a crypto-technical sense.

6. Conclusions

This paper has shown that, as with single word findings, there are considerable differences in collocational use across different academic domains, at least as far as our data from the two domains Maths and LAS are concerned. However, on top of the dimensions differentiating single word use between domains, there are also several other dimensions characteristic of collocational use that need to be taken into account in order to fully map collocational use in academic language and across academic domains. In particular, as well as the dimension covering differences between general, academic and domain-specific collocations, such analyses should also examine the lexical/grammatical collocational dimension, different collocational structures within the lexical/grammatical dimension, and nested collocations.

We have suggested possible avenues for solving the many problems inherent in analysing collocations from different domains, demonstrating what we feel is a robust method to identify collocations and the various dimensions which contribute to an informant's overall collocational use. This method will hopefully be reinforced with the inclusion of IT and other academic domains in our future research.

References

Ackermann, K., & Chen, Y. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes* (Impact Factor: 0.8) *12*(4), 235–247. doi:10.1016/j.jeap.2013.08.002

AILA World Congress 2014, 10–15 August 2014.

Ananiadou, S., & McNaught, J. (1995). Terms are not alone: Term choice and choice terms. *Journal of Aslib Proceedings, 47*(2), 47–60. doi:10.1108/eb051381

Barfield, A., & Gyllstad, H. (2009). Introduction: Researching L2 collocation knowledge and development. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 1–20). Basingstoke: Palgrave Macmillan.

Chung, T. & Nation, I.S.P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language, 15*(2), 103–116. doi:10.1016/j.system.2003.11.008

Chung, T. & Nation, I.S.P. (2004). Identifying technical vocabulary. *System, 32*(2), 251–263.

Complexity and Idiomaticity, Stockholm University, June 2012.

Coxhead, A. (2000). A New Academic Wordlist. *TESOL Quarterly, 34*(2), 213–238. doi:10.2307/3587951

Dang, T. (2016). Investigating vocabulary in academic spoken English: Corpora, teachers, and learners, PhD Thesis, Victoria University of Wellington.

EIE, The Copenhagen conference (19–21 April 2013): The English Language in Europe in Teaching in European Higher Education.

Frantzi, K.T. & Ananiadou, S. (1996). Extracting nested collocations. In Proceedings of the 16th International Conference on Computational Linguistics, 1996. COLING '96, pp. 41–46. Association for Computational Linguistics.

Fraser, S. (2006). The nature and role of specialized vocabulary: What do ESP teachers and learners need to know? *Hiroshima University Scholarly Journals, 2005*, 63–75.

Granger, S. & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds,), *Phraseology: An interdisciplinary perspective* (pp. 27–49). Amsterdam: John Benjamin.

Haag, N., Heppt, B., Stanat, P., Kuhl, P., & Pant, H.A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction,* 28, 24–34. doi:10.1016/j.learninstruc.2013.04.001

Handl, S. (2009). Towards collocational webs for presenting collocations in learners' dictionaries. In A. Barfield & H. Gyllstad (pp. 69–85). Multiple interpretations. Basingstoke: Palgrave Macmillan, 2009.

Henriksen, B. (2013). Research on L2 learners' collocational use and development – a progress report. In Bardel, Camilla; Laufer, Batia & Lindqvist, Christina (Eds.), *L2 vocabulary acquisition, knowledge and use. New perspectives on assessment and corpus analysis*. Eurosla Monographs Series, 2. EUROSLA.

Herbel-Eisenmann, B.A. (2002). Using student contributions and multiple representations to develop mathematical language. *Mathematics Teaching in the Middle School,* 8(2), 100.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, 19(1), 24–44. doi:10.1093/applin/19.1.24

Hwang, K. & Nation, P. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System, 23*(1), 35–41. doi:10.1016/0346-251X(94)00050-G

Laufer, B. & Waldman, T. (2011). Verb-noun collocations in second-language writing: A corpus analysis of learners' English. *Language Learning, 61*(2), 647–672. doi:10.1111/j.1467-9922.2010.00621.x

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299–320. doi:10.1093/applin/ams010

Méndez Cendón, B. (2004). *Medical language collocations: The case of the verb perform*. A New Spectrum of Translation Studies. Valladolid: Universidad de Valladolid, pp. 195–208.

Menon, S. & Mukandan, J. (2010). Analysing collocational patterns of semi-technical words in science textbooks. *The Social Science & Humanities, 18*(2), 241–258.

Moon, R. (1997). Vocabulary connections: Multi-word items in English. In N. Schmitt & M. McCarthy (Eds), *Vocabulary description, acquisition and pedagogy* (pp. 40–63). Cambridge: Cambridge University Press.

Nation, I.S.P. (1990). *Teaching and learning vocabulary.* New York: Newbury House.

Nation, I.S.P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139524759

Nattinger, J.R. & DeCarrico, J.S. (1992). *Lexical phrases and language teaching.* Oxford: Oxford University Press.

Nesselhauf, N. (2005). Collocations in a learner corpus. *Studies in Corpus Linguistics, 14*. doi:10.1075/scl.14. John Benjamins

PhD Applied Linguistics (Lexical studies) annual conference, 10-13th of March, 2014, Cardiff, Wales.

Pulverness, A. (2007). Review of English collocations in use. *ELT Journal, 61*(2). doi:10.1093/elt/ccm014

SDU SELC Conference, Odense 2013.

Westbrook, P. (2015). Talk about mouth speculums: Collocational use and spoken fluency in non-native English-speaking university lecturers. *Studies in Parallel Language Use, C8*, Copenhagen Studies in Bilingualism. University of Copenhagen.

Xue, G., & Nation, I.S.P. (1984). A university word list. Language Learning and Communication*,* *3*(2), 215–229.