OSLa
Fjeld, Hagen, Henriksen, Johansson, Olsen & Prentice (eds.) *Academic Language in a Nordic Setting – Linguistic and Educational Perspectives*, Oslo Studies in Language 9(3), 2017. 9–27. (ISSN 1890-9639)
http://www.journals.uio.no/osla

ACADEMIC PHRASEOLOGY A KEY INGREDIENT IN SUCCESSFUL L2 ACADEMIC LITERACY

SYLVIANE GRANGER

ABSTRACT

One of the weaknesses of most current academic word lists is that they fail to do justice to the large stock of multiword units that are typical of academic language. The objective of this chapter is to raise awareness of the importance of phrasal academic vocabulary. After a brief critical survey of three recently compiled phrasal academic lists, the chapter highlights the potential contribution of learner corpus data to identifying the most useful units for teaching purposes. The approach is illustrated with a case study of phrasal metadiscourse based on corpora of novice and expert native writing, and subcorpora from the *International Corpus of Learner English* representing L2 writers from six different mother tongue backgrounds.

[1] INTRODUCTION

Academic vocabulary can be roughly defined as the words and phrases that are typically used in academic contexts. While acquiring these lexical items is essential to the academic achievement of both native and non-native (L2) students, they represent a particularly significant hurdle for L2 users, who have to understand and produce academic language in a language that is not their own.

Academic vocabulary is commonly subdivided into discipline-specific academic words, i.e. technical words that describe content knowledge (e.g. *malignant, biopsy* or *incubate* in medicine), and general academic words that occur across content areas and are used to refer to activities typical of academic work and to structure discourse (e.g. *despite, conclusion* or *categorise*). Both types of word are challenging for L2 learners, but it is the cross-disciplinary ones that pose the greatest difficulties. One of the reasons is that these words 'are supportive of but not central to the topics of the texts in which they occur' (Coxhead 2000: 214). As a result, they are not particularly salient and tend to pass unnoticed. To help teachers give these words the pedagogical attention

they require, several lists of cross-disciplinary academic words have been compiled. The first to appear was Coxhead's (2000) Academic Word List (AWL), which quickly met with great success and is still the most widely used today. However, as acknowledged by Coxhead (2008) herself, '[o]ne of the challenges of the AWL is that it was released solely as a list of individual words and their families, with no indication of the context and patterning in which these words occurred'. The same weakness is pointed out by Gardner & Davies (2013), the compilers of the Academic Vocabulary List, who conclude that 'more needs to be done in the future to identify core multiword academic vocabulary' (p. 325). A first step in that direction was made by Paquot (2010), whose Academic Keyword List contains a number of multiword adverbs, prepositions and conjunctions (e.g. *as well as, according to, as opposed to, as to, contrary to, in favour of, rather than, for example, whether or not*), though these represent but a small proportion (c. 3%) of the whole list. Another initiative designed to 'phrase up' single-word vocabulary lists is that of Martinez & Schmitt (2012), who, using a mixture of automatic extraction and manual vetting by judges with language testing and teaching experience, put together a Phrasal Expressions List, which contains 505 frequent non-transparent multiword expressions in English, specially intended for receptive use.

In recent years, there have been several initiatives to produce lists of word combinations typical of academic discourse (Durrant 2009, Simpson-Vlach & Ellis 2010, Ackermann & Chen 2013). The main objective of this chapter is to take a critical look at these lists and the methods used to generate them (Section 2) and to describe the potential contribution of learner corpora to identifying the most useful units (Section 3). In Section 4, the corpus-based approach to academic phraseology is illustrated by means of a case study of phrasal metadiscourse. Section 5 offers some conclusions and avenues for future research.

[2] PHRASAL ACADEMIC LISTS

Phrasal academic lists are lists of phrasemes (Mel'čuk 1998), i.e. word combinations displaying some degree of selectional restriction that are deemed to offer high learning and teaching potential. Those that have been compiled to date target two different types of word combination: collocations, i.e. pairs of words that co-occur in a short span of text more often than predicted by chance (Sinclair 1991) and are identified by means of statistical tests such as Mutual Information (MI), and lexical bundles, i.e. sequences of contiguous words that recur in a particular register (Biber et al. 1999) and are extracted using the n-gram technique. One characteristic shared by the two types of unit is that they are usually not very difficult to understand, but are very difficult to get right in

productive tasks.

The lists assembled by Durrant (2009) and Ackermann & Chen (A&C) (2013) contain typical English for Academic Purposes (EAP) collocations extracted from large corpora of native or expert academic texts. Although they were compiled with the same objective – that of providing a pedagogically useful resource –, the two lists are quite different, not only in size (1,000 pairs for Durrant vs. 2,468 for A&C) but also in the quality of the units. Durrant's list contains a majority (76%) of grammatical collocations, i.e. containing at least one grammatical word, such as *consistent with, as shown, suggest that*, while A&C's Academic Collocation List is limited to lexical collocations made up of two lexical words (see examples in Table 1). Several other factors account for discrepancies between the two lists, among them the following four: (1) the corpora used differ in size and academic genres covered; (2) A&C exclude collocation pairs that contain words from West's (1953) General Service List (top 2,000 words), while Durrant includes them; (3) the collocation status of the word pairs relies on different quantitative criteria (e.g. MI of minimum 4 for Durrant and minimum 3 for A&C); (4) Durrant relies solely on automatic extraction, while A&C's approach makes use of both automatic extraction and manual screening by linguists (e.g. to exclude highly fixed units[1] such as *collective bargaining*) and language practitioners (to select the pedagogically most relevant units).

Verb	Adverb	Adjective	Noun
differ	significantly	causal	link
expand	rapidly	conflicting	interests
explore	further	crucial	factor
increase	dramatically	final	stage
vary	widely	further	information

TABLE 1: Excerpt from Ackerman & Chen's (2013) Academic Collocation List

Rather than collocations, the Academic Formulas List (AFL) put together by Simpson-Vlach & Ellis (2010) comprises lexical bundles, i.e. uninterrupted sequences of 3-5 words extracted from academic corpora. The final list contains 600 formulaic sequences, subdivided into three 200-formula lists (core, spoken and written). Table 2 shows the top 15 spoken and written sequences in the AFL. The list was obtained using a mixture of automatic and manual extraction procedures. The first step consisted in the automatic extraction of sequences

[1] The degree of fixedness was determined by consulting several online dictionaries to see whether the word combinations were listed as independent entries.

that were more frequent in academic than non-academic texts. Conscious that 'long lists of highly frequent expressions are of minimal use to instructors who must make decisions about what content to draw students' attention to for maximum benefit within limited classroom time', Simpson-Vlach & Ellis (2010: 490) submitted a sample of the data to experienced teachers, who were asked to assess their usefulness for teaching purposes. On that basis they created a metric, called the Formula Teaching Worth, which reflects the statistical correlation between teacher intuition, on the one hand, and phrase frequency and MI score, on the other, and used it to draw up the final list. To enhance the usefulness of the list for teachers, the authors then grouped the sequences into discourse-pragmatic categories. For example, bundles such as *a high degree*, *a large number of*, *a wide range of* were classified in the category of 'quantity specification', while *it might be*, *is likely to*, *in a sense* were classified as 'hedges'.

Spoken AFL	**Written AFL**
be able to	on the other hand
blah blah blah	due to the fact that
this is the	on the other hand the
you know what I mean	it should be noted
you can see	it is not possible to
trying to figure out	a wide range of
a little bit about	there are a number of
does that make sense	in such a way that
you know what	take into account the
the University of Michigan	as can be seen
for those of you who	it is clear that
do you want me to	take into account
thank you very much	can be used to
look at the	in this paper we
we're gonna talk about	are likely to

TABLE 2: Top 15 lexical bundles in Simpson-Vlach & Ellis's (2010) Spoken and Written Academic Formulas List

These initiatives are promising, but they are only a beginning. The discrepancy between the lists shows that there are still a considerable number of issues to be addressed in future research, including the following:

(i) What quantitative criteria should best be used to extract the units? On the basis of what statistical tests and with what thresholds?

(ii) Should some types of unit be excluded from the lists and, if so, which ones and on the basis of what criteria?

(iii) Should the automatically derived lists be filtered by teachers and, if so, on the basis of what criteria?

Whatever the method used, the lists generated tend to be quite long, and it is not realistic to expect teachers to have the time or inclination to give all of them the same degree of pedagogical attention. As argued in the following section, one particularly helpful way of identifying the most useful combinations is to resort to learner corpus data.

[3] INSIGHTS FROM LEARNER CORPORA

To extract academic phrases, researchers have so far relied solely on native or expert corpora. This method is clearly essential to identifying the most typical native-like units. However, it provides no information on the basis of which to select the most useful units for specific learning/teaching contexts. One essential feature, namely the degree of difficulty for a particular learner population, is completely disregarded. To bring this variable into the picture, it is necessary to complement corpora of native or expert language with learner corpus data, i.e. electronic collections of academic writing and speech by L2 learners.

Learner corpora are a relatively new resource that is enjoying growing interest from the language education community at large, and specialists in academic writing in particular. One of the advantages such corpora offer is that they tend to be quite large and therefore provide numerous authentic examples of learners' difficulties in context. In addition, as the data are in electronic format, they can be explored automatically with the help of powerful software tools, thereby offering insights that would be inaccessible to manual analysis.

Using the methodology referred to as 'contrastive interlanguage analysis' (Granger 2015), it is possible to extract entirely automatically multiword units that are used significantly more frequently (overuse) or less frequently (underuse) in learner corpora than in comparable native or expert texts, as well as to compare frequency of use across learner populations (e.g. Swedish vs. Spanish learners). Misuse can also be detected, such as learners' use of *in the other hand* or *on the other side* instead of, or in addition to, the nativelike connector *on the other hand*.

This powerful methodology has been used in a large number of studies, some focused on collocations, others on lexical bundles, rarely on both types of unit. For reasons of space it is not possible here to sum up the main trends emerging from these studies (for recent surveys, see Paquot & Granger 2012

and Ebeling & Hasselgård 2015). However, one trend deserves special mention because it is consistent across the studies, viz. the high degree of transfer from the learner's mother tongue (L1). For example, in her study of the use of verb-noun combinations by advanced German-speaking learners, Nesselhauf (2005) observes that some 50% of the inappropriate collocations are probably transfer-related. An even higher percentage (67%) is given in Wanner et al.'s (2013) study of miscollocations by Spanish learners. L1 transfer is also at play in the use of lexical bundles. In her study of lexical bundles in English texts written by French-speaking learners, Paquot (2013) finds different types of idiosyncratic use which can be traced back to French. She attributes the ease with which lexical bundles can be transferred to the fact that they are semantically and syntactically compositional and therefore not sufficiently salient to attract learners' attention. This finding has important pedagogical implications: it means that in order to ensure maximum efficiency, the pedagogical attention given to certain phraseological units should not be the same in the case of all learners but adapted to the attested needs of particular learner populations.

The flurry of publications focused on phraseology in learner corpus research demonstrates the benefits of an approach that extends the corpus base to incorporate learner corpus data in addition to the traditionally used native/expert data. In the next section this approach is illustrated by means of a case study on phrasal metadiscourse.

[4] CASE STUDY: PHRASAL METADISCOURSE

Metadiscourse, i.e. the use of language to 'organise texts, engage readers and signal attitudes to the material and the audience' (Hyland 2015: 1), is often expressed by sequences of words rather than single words and is therefore a particularly suitable area of language in which to investigate the phrasemes used by learners when they write academic texts. This section illustrates how the combined use of native and learner corpus data can help uncover differences in the preferred phraseological patterning of one particular metadiscursive word – the noun *conclusion* (Section 4.1) – and in the use of four-word metadiscursive sequences (Section 4.2). The first analysis is corpus-based, i.e. it starts from a word that is assumed to be worth investigating, and corpus tools and methods are used to retrieve all its occurrences and explore its phraseology. The second is corpus-driven, i.e. no assumption is made initially as to the linguistic items to be investigated; rather, it is the corpus tools and methods that automatically generate lists of potentially interesting phraseological sequences.

The methodology used for the analysis involves the two branches of contrastive interlanguage analysis, i.e. comparison of learner varieties with one or

more reference varieties, and comparison between learner varieties. The learner corpus data come from the *International Corpus of Learner English* (ICLE) (Granger et al. 2009), which contains essays written by higher intermediate to advanced learners from sixteen mother tongue backgrounds. The subcorpus used contains argumentative essays[2] from learner populations representing six mother tongue backgrounds, i.e. French (FR), German (GE), Italian (IT), Norwegian (NO), Spanish (SP) and Swedish (SW). The composition of the learner corpus allows comparisons not only between individual L2 populations, e.g. French-speaking vs. German-speaking learners, but also between two groups of L2 populations, i.e. Romance vs. Germanic.

Two reference corpora were used for the comparison with the L2 data. Both are native English corpora, but they represent different degrees of academic maturity. The *Louvain Corpus of Native English Essays* (LOCNESS) is a corpus of argumentative writing by university students and therefore represents novice native writing. Its main advantage is that it is fully comparable with the ICLE in terms of genre. The other native English corpus is the academic section of the *British National Corpus* (BNC), *Baby.*[3] Made up of academic texts from periodicals and books, it represents professional academic writing. Having two reference points will make it possible to assess whether, as stated by Römer (2009), L2 learners tend to behave similarly to novice native writers or whether, as argued by Gilquin et al. (2007), they display features that set them distinctly apart from native writers, whether novice or professional. The size of the eight subcorpora is shown in Table 3.

Corpus	No. of words
ICLE-FR	160,245
ICLE-GE	228,180
ICLE-IT	199,001
ICLE-NO	207,230
ICLE-SP	127,051
ICLE-SW	162,216
LOCNESS	327,807
BNC-Acad	1,027,550

TABLE 3: Size of the eight subcorpora

[2] The literary essays in the ICLE were excluded from the data set.
[3] http://www.natcorp.ox.ac.uk/archive/oldBabyDocs/baby-des.html

4.1 *Corpus-based approach:* conclusion

The simple extraction of all the occurrences of *conclusion* from the eight subcorpora already reveals some interesting results. As can be seen in Figure 1, there are quite striking differences between the Romance learner populations (SP, FR and IT), which are characterised by heavy use of *conclusion*, and the Germanic learner populations (NO, SW and GE), whose frequency of use is much closer to that displayed by novice and professional native speakers (in black in the Figure).

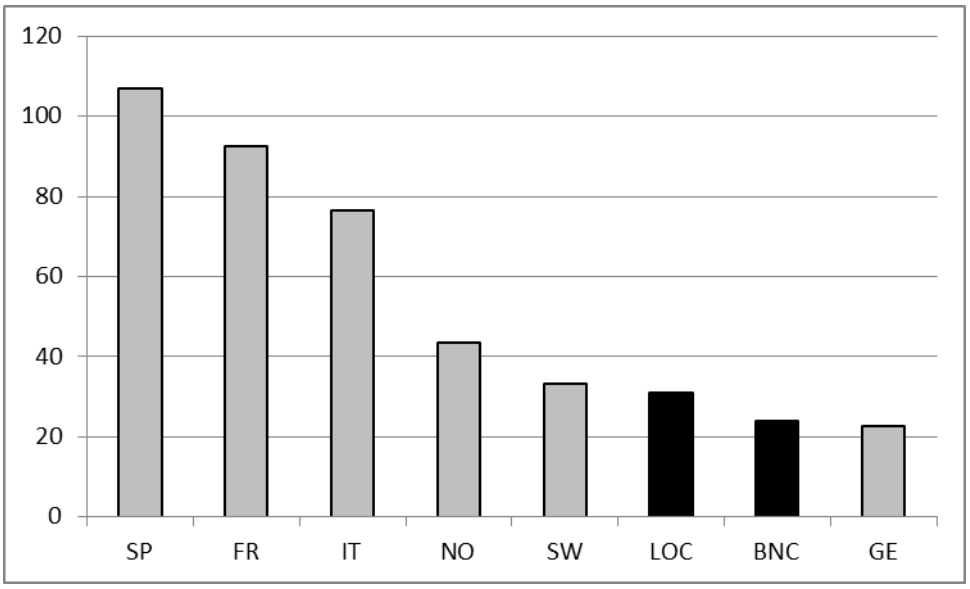

FIGURE 1: Relative frequency of the word *conclusion* in the eight subcorpora (/200,000 words)

Over and above this purely quantitative difference, the analysis also reveals some interesting qualitative findings. The noun *conclusion* can be used in two different ways: as part of a noun phrase functioning as subject or object (e.g. *a conclusion that can be drawn*) and as part of an adverbial connector, usually used at the beginning of the sentence, followed by a comma (e.g. *In conclusion, this study shows*). A close scan of the concordance lines shows that the proportion of connector vs. non-connector use varies widely across the L2 subcorpora. As shown in Figure 2, while some learner populations mainly use *conclusion* as an adverbial connector (91% in IT), others rarely use it in that way (only 15% in GE).

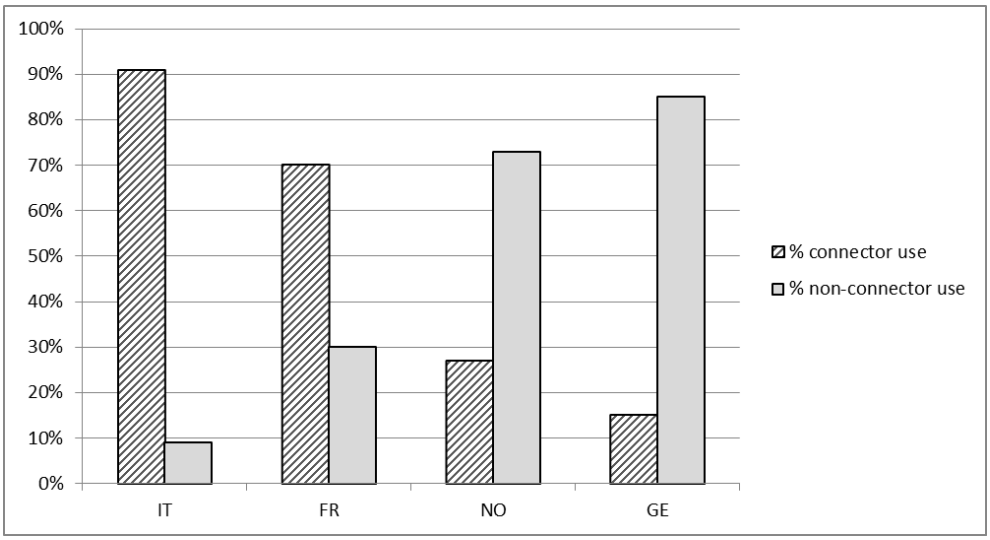

FIGURE 2: Connector vs. non-connector use of *conclusion* (percentage of use)

In addition, close scrutiny of the occurrences reveals three types of difficulty experienced by learners when using the word *conclusion.* First, learners regularly make use of atypical adverbial connectors such as *as a conclusion* or *as conclusion* instead of the usual *in conclusion* (see example 1). In some cases the learner-idiosyncratic connectors are even more frequent than the standard connector. For example, in the FR subcorpus *as a conclusion* accounts for 73% of all the connector uses, as against only 27% for *in conclusion.* In a more extended study involving 11 learner populations of the ICLE, Paquot (2010) established that *as a conclusion* represents approximately 40% of the concluding phrasemes involving the noun *conclusion*. Findings of this type would have been missed by studies focused exclusively on the standard connector. The second difficulty concerns the use of *conclusion* as a verb argument. Instead of using the typical verbs (*come to/reach a conclusion, draw a conclusion, offer a conclusion,* etc.), learners often opt for atypical verbal collocates (*reach to* in example 2 and *make* in example 3).

(1) *As a conclusion*, we can say that the political and cultural unity... (FR)
(2) With this idea we *reach to* the conclusion that a chaos is continually dominating our world (SP)
(3) I let it be entirely up to you to *make* a conclusion (NO)
(4) *In conclusion*, the root of all evil is the choice of the individual (LOCNESS)
(5) *In conclusion, I want to point out that in my own view*, capital punishment... (GE)

(6) *As a conclusion I would only like to say that I think that* we are overreacting wildly... (SW)

As a conclusion, I am of the opinion that...
As a conclusion, I can say that...
As a conclusion, I would like to say that...
As a conclusion, I would say that...
As a conclusion, I would then say that...
As a conclusion, it can be said that...
As a conclusion, one can say that...
As a conclusion, one could say that...
As a conclusion, we can say that...
As a conclusion, we can see that...
As a conclusion, we could say that...
As a conclusion, we may ask...
As a conclusion, we may say that...
In conclusion, I would say that...
In conclusion, I would simply like to say that...
In conclusion, one could say that...

FIGURE 3: Extended metadiscursive sequences in ICLE-FR

Finally, the learners produce extended metadiscursive sequences made up of an adverbial connector containing the word *conclusion* (*as a conclusion* or *in conclusion*) and a stance bundle mostly with the verb *to say*, which add redundancy and verbosity to their texts. In ICLE-FR, 54% of the total number of occurrences of *conclusion* occur in this type of sequence and, as shown in Figure 3, the diversity of the sequences is very high.[4] In native writing, the concluding statement usually follows immediately after the connector (see example 4). Admittedly, extended metadiscursive sequences can also be found in native writing, but they occur in much smaller numbers and never display the kind of metadiscursive overkill to be found in some of the learner sequences (see examples 5 and 6).

4.2. Corpus-driven approach: four-word metadiscursive bundles

In a corpus-driven approach, no a priori assumptions are made about the specific linguistic forms to be investigated. The first stage of the analysis is fully automatic: it involves extraction of all the four-word sequences from the eight subcorpora. This stage is followed by manual selection of metadiscursive sequences, which either have an organisational function, i.e. reflect relationships between prior and coming discourse (e.g. *on the other hand*), or express stance, i.e. attitude or assessment of certainty (e.g. *it is true that*) (Biber et al.

[4] See Paquot (2010: 160-161) for a discussion of this phenomenon based on Gledhill's (2000) notion of 'phraseological cascade'.

2004: 384).

As the aim was to compare the sequences preferred by each population – those which, based on Hasselgren's (1994) metaphor, are referred to by Ellis (2012) as 'phrasal teddy bears' –, the top 20 sequences were selected from each corpus, resulting in a general list of 81 metadiscursive bundle types (see the appendix for the full list). Before turning to some of the results, some caveats are in order. First, this study is exploratory; its main purpose is to illustrate a methodological approach rather than to establish hard facts. The sequences identified as metadiscursive are in fact potentially metadiscursive, as they have not been subjected to manual disambiguation in context. Second, the corpora are not very large, which may reduce the representativeness of the results. Third, dispersion within the different subcorpora has not been measured. However, the study suggests some interesting – sometimes intriguing – differences between learner populations and native writers which can be used as starting points for more extended studies.

A comparison of the top 20 sequences in each corpus reveals a much higher degree of recurrence in the learner data, the only exception being the German learner subcorpus (see Figure 4). Interestingly, the novice and professional native corpora have a very similar degree of recurrence, which is roughly half that displayed by most of the learner data.

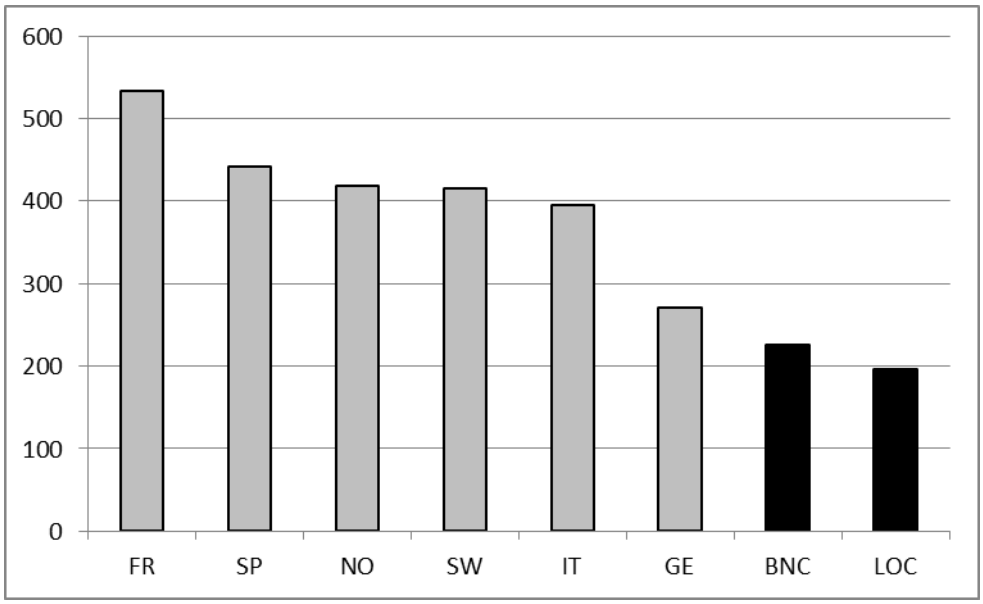

FIGURE 4: Frequency of the top 20 bundles (tokens/200,000 words)

This quantitative difference comes with a range of qualitative differences. As can be seen in Table 4, which lists the top 20 bundles in each corpus, the bundles in the learner data tend to be more clausal (*we can say that, it is important to*) and more involved, i.e. "reflecting interpersonal interaction and the involved expression of personal feelings and concerns" (Biber et al. 1998: 150). The sequences containing the first person pronouns (*I* and *we*) and the possessive determiner *my* (in bold in the table) are a distinctive feature of learner corpora. Both the novice and the professional native texts tend to contain more phrasal bundles, i.e. noun- rather than verb-based bundles (*as a result of, on the basis of*) and fewer involved bundles, and generally display a clear tendency towards impersonal stance (*it is obvious that, it is important to*).

FR	SP	IT	NO	SW	GE	LOCNESS	BNC
On the other hand	On the other hand	On the other hand	On the other hand	When it comes to	On the other hand	On the other hand	On the other hand
At the same time	At the same time	It is clear that	When it comes to	On the other hand	At the same time	At the same time	In terms of the
As far as the	In the case of	As a matter of	**I do not think**	At the same time	**I would like to**	As a result of	In the case of
On the one hand	**I would like to**	It is true that	At the same time	**I think it is**	On the one hand	It is obvious that	On the basis of
I would like to	That is to say	At the same time	It is important to	**I do not think**	Last but not least	In the case of	As a result of
It is true that	It is true that	**I think that the**	**I would say that**	**I would like to**	As a matter of	As well as the	It is possible to
As a matter of	As a result of	**But I think that**	Some people say that	It is important to	In the same way	Due to the fact	In the context of
I do not think	**My point of view**	In the same way	**I think it is**	It is hard to	**I must admit that**	When it comes to	At the same time
I would say that	In the same way	In the case of	**I would like to**	It is impossible to	It is true that	It is important for	It is important to
It is obvious that	It is very difficult	That is to say	**In this essay I**	In the same way	That is to say	It is important to	It is clear that
That is to say	On the one hand	**I do not think**	People say that in	**In this essay I**	**As far as I**	It is hard to	As well as the
I think that the	The fact is that	**In my opinion the**	It is possible to	As a result of	**I am convinced that**	A good example of	It is difficult to
We can say that	**We can say that**	Is the result of	There is no doubt	**I will try to**	**I think it is**	In an attempt to	**As we have seen**
The problem is that	As a matter of	**And I think that**	As a result of	The problem is that	It seems to be	It is clear that	It is necessary to
As a conclusion I	It is said that	First of all the	**I believe that the**	Due to the fact	On the other side	On the issue of	It has been suggested
I think it is	It is obvious that	In fact it is	It is a fact	**I think that we**	As well as the	**I would like to**	In the same way
In the same way	**In my opinion it**	It is important to	It is easy to	**I think that the**	**In my opinion this**	An example of this	**As we shall see**
Take the example of	**In my opinion the**	It is not possible	It is hard to	It is important that	Due to the fact	The purpose of this	It may be that
As well as the	**As we all know**	It would be better	**We do not know**	It is difficult to	**I am sure that**	**I have come to**	It can be seen
But I think that	**As we can see**	Due to the fact	**I am going to**	It is true that	**I do not think**	The problem is that	**In this chapter we**

TABLE 4: Top 20 metadiscursive bundles in the learner and native subcorpora

As illustrated in Table 5, there are also marked differences in the frequency of use of individual bundles across learner and native speaker populations. The bundle *as a conclusion I* is a hallmark of French-speaking learners, while *we can say that* is a typical Romance use, hardly found in any of the other corpora. *On the other hand* is a favourite with all learners, but the French- and Spanish-speaking learners make particularly high use of this connector. The sequence *when it comes to* is a Nordic phrasal teddy bear, much used by Norwegian and Swedish learners. In a study based on novice L1 and L2 linguistics research papers, Hasselgård (forthcoming) notes the same overuse in Norwegian learners and attributes it to the phrase *når det gjelder,* which is formally and functionally similar to *when it comes to* and is frequently used in Norwegian academic texts.[5]

Differences in preferential use of metadiscursive bundles should come as no surprise, as languages, not only genetically unrelated ones such as English and Arabic (Sultan 2011), but also closely related ones such as English and French (Granger 2014), differ markedly in their use of metadiscourse. Since sequences such as those investigated in this study are relatively inconspicuous, they tend to be transferred by learners into the target language in their entirety.

Metadiscursive bundle	FR	IT	SP	NO	SW	GE	LOC	BNC
on the other hand	84	54	94	57	49	50	23	25
when it comes to	1	1	3	40	53	6	10	0
as a conclusion I	16	0	2	3	2	0	0	0
we can say that	21	10	16	0	1	2	0	2

TABLE 5: Relative frequency (tokens/200,000 words) of four bundles across learner and native populations

A hierarchical cluster analysis (HCA) was also performed in order to obtain an overview of how the various populations compare in terms of their choice of metadiscursive bundles. HCA is an exploratory technique that computes a measure of similarity/dissimilarity between groups and provides a general overview of the dataset by means of a dendrogram. More specifically, the input data for the present HCA is based on the distribution (as a percentage) of the 81 metadiscursive bundles within each corpus. The aim was to establish whether (1) the novice native writers would tend to cluster with the novice non-native writers or with the professional native writers; and (2) whether the Romance and Germanic groups would tend to form distinct clusters. As the dendrogram

[5] The same explanation holds for Swedish, which has a similar phrase (*när det gäller*) that is even more frequent than its Norwegian equivalent (information gratefully received from Hilde Hasselgård).

in Figure 5 illustrates, native writers form their own cluster whatever their degree of academic maturity, which suggests that, in the case of this particular linguistic phenomenon at least, the notion of 'nativeness' is still a valid variable, over and above that of 'noviceness'. The results for the learner writers are more mixed: while the Romance learners cluster together, the Germanic learners are split: the German-speaking learners cluster with the Romance learners, while the Nordic group – Norwegian and Swedish – form a separate cluster closer to the native speakers.

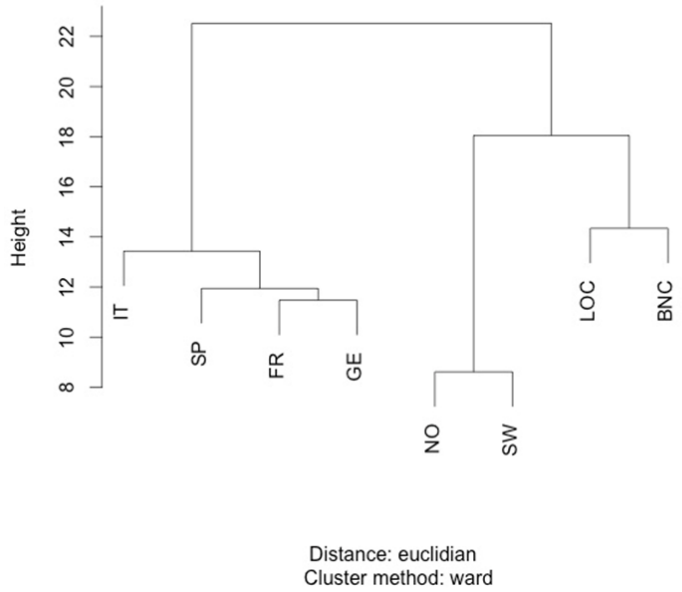

Figure 5: Hierarchical cluster analysis of the 81 metadiscursive bundles

[5] CONCLUSION

Phraseology is now recognised as a major component in general L2 learning and teaching. In the specialised field of academic literacy, however, the phraseological dimension has yet to establish itself as a core facet. The compilation of phrasal academic lists is a first sign that researchers are becoming aware of this shortcoming and are keen to provide resources to help both learners and teachers. As our analysis of three recently compiled lists of academic phrasemes has shown, researchers have experimented with all kinds of methods to generate the most useful lists. This is a necessary and indeed healthy step, but the results can be quite disconcerting for language practitioners who are presented with different lists and may not know which one to choose. In

fact, as each of the lists is based on its own set of selection criteria, the units they contain can be quite different and, as a result, should be seen as complementary rather than conflicting. Having lists is not enough, however. I fully agree with Simpson-Vlach & Ellis (2010: 510) that it is essential to organise the lists so as to turn them 'into something that might usefully inform curriculum or language testing materials'. These two authors pave the way by subdividing the bundles into the three categories of spoken, written and core bundles and, more importantly, by placing them in useful functional categories.

One voice that is rarely heard, however, is that of the learners themselves. As I hope to have demonstrated, learner corpus data is an invaluable resource for identifying the units that are likely to pose problems – whether in terms of misuse, overuse or underuse – for learners in general or for specific learner populations. Learner corpora can be explored in two ways: by searching for a given word that is known or suspected to be problematic, or by letting the corpus speak, i.e. employing computational methods to extract multiword units typically used by learners. Though small-scale and largely exploratory, the two studies I have carried out to illustrate these two methods have revealed interesting aspects of the learner phrasicon. In particular, the results highlight marked differences between L2 learners and native writers, be they novice or professional. The study also shows that L2 learners should not be considered a homogeneous group: while some of the learner-idiosyncratic features are generic, i.e. shared by all the learner populations, most are L1-specific and require differentiated pedagogical attention.

The key issue that is left unaddressed by the present study is how to incorporate all this information into teaching practice. Like Coxhead & Byrd (2012: 19), I am convinced that '[t]eaching applications using data sets such as we present here must be mediated. Taking raw data and linguistic techniques into the classroom requires a great deal of care'. I also agree that 'corpus-based dictionaries and grammars are a wise approach at this time', but I would go one step further and specify the format in which these dictionaries and grammars should be presented. In my view, the only way to guarantee the kind of flexibility that is needed to adapt the resource to different learner groups is to design customisable web-based environments which learners can turn to when they write academic texts. The Louvain English for Academic Purposes Dictionary (Granger & Paquot 2015), a web-based dictionary-cum-writing aid, which provides a wealth of information on the phraseology of academic words (collocations and lexical bundles), as well as generic and L1-specific warnings against recurrent errors, is a first step in that direction and will hopefully inspire further research in this field.

REFERENCES

Ackermann, Kirsten & Yu-Hua Chen. 2013. Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes* 12(4), 235-247.

Biber, Doug, Susan Conrad & Randi Reppen. 1998. *Corpus Linguistics. Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Biber, Doug, Susan Conrad & Viviana Cortes. 2004. If you Look at ... Lexical Bundles in University Lectures and Textbooks. *Applied Linguistics* 25, 371-405.

Biber, Doug, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Coxhead, Averil. 2000. A new academic word list. *TESOL Quarterly* 34(2), 213-238.

Coxhead, Averil. 2008. Phraseology and English for academic purposes. In Fanny Meunier & Sylviane Granger (eds.), *Phraseology in Foreign Language Learning and Teaching,* 149-161. Amsterdam: Benjamins.

Coxhead, Averil & Patricia Byrd. 2012. Collocations and Academic Word List: The strong, the weak and the lonely. In Isabel Moskowich & Begoña Crespo (eds.), *Encoding the Past, Decoding the Future: Corpora in the 21st Century,* 1-20. Cambridge: Cambridge Scholars Publishing.

Durrant, Philip. 2009. Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes* 28, 157-169.

Ebeling, Signe Oksefjell & Hilde Hasselgård. 2015. Learner corpora and phraseology. In Sylviane Granger, Gaëtanelle Gilquin & Fanny Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research,* 207-229. Cambridge University Press.

Ellis, Nick. 2012. Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics* 32, 17-44.

Gardner, Dee & Mark Davies. 2013. A New Academic Vocabulary List. *Applied Linguistics* 35, 305-327.

Gilquin, Gaëtanelle, Sylviane Granger & Magali Paquot. 2007. Learner corpora: the missing link in EAP pedagogy. *Journal of English for Academic Purposes*

6(4), 319-335.

Gledhill, Chris. 2000. *Collocations in Science Writing*. Language in Performance 22. Tuebingen: Gunter Narr Verlag.

Granger, Sylviane. 2014. A lexical bundle approach to comparing languages: Stems in English and French. *Languages in Contrast* 14:1, 58-72.

Granger, Sylviane. 2015. Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1(1), 7-24.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier & Magali Paquot. 2009. *The International Corpus of Learner English. Handbook and CD-ROM. Version 2*. Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, Sylviane & Magali Paquot. 2015. Electronic lexicography goes local: Design and structures of a needs-driven online academic writing aid. *Lexicographica - International Annual for Lexicography* 31(1), 118-141.

Hasselgård, Hilde. Forthcoming. Phraseological teddy bears: frequent lexical bundles in academic writing by Norwegian learners and native speakers of English. To appear in Michaela Mahlberg and Viola Wiegand (eds), *Corpus Linguistics, Context and Culture*. Berlin: De Gruyter Mouton.

Hasselgren, Angela 1994. Lexical teddy bears and advanced learners: a study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics* 4(2), 237–258.

Hyland, Ken. 2005. *Metadiscourse*. London: Continuum.

Martinez, Ron & Norbert Schmitt. 2012. A phrasal expressions list. *Applied Linguistics* 33(3), 299-320.

Mel'čuk, Igor. 1998. Collocations and lexical functions. In Anthony Cowie (ed.). *Phraseology. Theory, Analysis, and Applications*, 23–53. Oxford: Oxford University Press.

Nesselhauf, Nadja. 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

Paquot, Magali. 2010. *Academic Vocabulary in Learner Writing. Form Extraction to Analysis*. London & New York: Continuum.

Paquot, Magali. 2013. Lexical bundles and L1 transfer effects. *International Jour-*

nal of Corpus Linguistics 18(3), 391-417.

Paquot, Magali & Sylviane Granger. 2012. Formulaic language in learner corpora. *Annual Review of Applied Linguistics* 32, 130-149.

Römer, Ute. 2009. English in Academia: Does nativeness matter? *Anglistik: International Journal of English Studies* 20(2), 89-100.

Simpson-Vlach, Rita & Nick Ellis. 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics* 31(4), 487-512.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sultan, Abbas. 2011. A Contrastive Study of Metadiscourse in English and Arabic Linguistics Research Articles. *Acta Linguistica* 5(1), 28-41.

Wanner, Leo, Margarita Alonso Ramos, Orsolya Vincze, Rogelio Nazar, Gabriela Ferraro, Estela Mosqueira & Sabela Prieto. 2013. Annotation of collocations in a learner corpus for building a learning environment. In Sylviane Granger, Gaëtanelle Gilquin & Fanny Meunier (eds.), *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead,* 493-503. Presses universitaires de Louvain: Louvain-la-Neuve.

West, Michael. 1953. *A General Service List of English Words*. London: Longman.

APPENDIX: LIST OF 81 METADISCURSIVE SEQUENCES

a good example of	I do not think	it can be seen	last but not least
an example of this	I have come to	it has been suggested	my point of view
and I think that	I must admit that	it is a fact	on the basis of
as a conclusion I	I think it is	it is clear that	on the issue of
as a matter of	I think that the	it is difficult to	on the one hand
as a result of	I think that we	it is easy to	on the other hand
as far as I	I will try to	it is hard to	on the other side
as far as the	I would like to	it is important for	people say that in
as we all know	I would say that	it is important that	some people say that
as we can see	in an attempt to	it is important to	take the example of
as we have seen	in fact it is	it is impossible to	that is to say
as we shall see	in my opinion it	it is necessary to	the purpose of this
as well as the	in my opinion the	it is not possible	the fact is that
at the same time	in my opinion this	it is obvious that	the problem is that
but I think that	in terms of the	it is possible to	there is no doubt
due to the fact	in the case of	it is said that	we can say that
first of all the	in the context of	it is true that	we do not know
I am convinced that	in the same way	it is very difficult	when it comes to
I am going to	in this chapter we	it may be that	
I am sure that	in this essay I	it seems to be	
I believe that the	is the result of	it would be better	

CONTACT

Sylviane Granger
Université catholique de Louvain
sylviane.granger@uclouvain.be