# Lexical Bundles in EMI lectures: An exploratory Study

*Alessandra Molino*

*Abstract*
Drawing on studies of the linguistic challenges posed by English-medium instruction (EMI) to lecturers, this paper investigates the use of lexical bundles in two small, specialised corpora (ca. 65,000 words each) of Physical Sciences and Engineering lectures delivered in English in Anglophone (the UK) and non-Anglophone (Italy) university settings. The goal is to offer initial insights into the recurrent structures, functions and distribution of lexical bundles in EMI lectures. These multi-word units are focused upon because they are an important component of the university teaching register and because they contribute to fluency. The results suggest that the EMI lectures in the corpus are less formulaic than comparable lectures from the UK, in terms of the number of bundle types and tokens. They are also more 'unconventional', since they contain more bundles not found in previous studies of academic lectures. Finally, bundles conveying pragmatic meanings, particularly vagueness markers and hedges, are absent from the EMI lectures, hinting at a higher level of formality.

*Keywords*: EMI lectures, lexical bundles, physical sciences and engineering, Turin EMI lecture corpus, BASE corpus.

## 1. Introduction

Globalisation and the internationalisation of higher education are promoting the increasing use of English as a Medium of Instruction (EMI) in European universities. Italian academia is no exception: universities are gradually embracing EMI, especially in the areas of the Physical Sciences, Engineering and Economics (Campagna and Pulcini 2014; Broggini and Costa 2017). However, the introduction of EMI programmes is not friction-free. Various studies have documented conflicting attitudes among stakeholders (e.g. Molino and Campagna 2014; Pulcini and Campagna 2015), highlighting the

role of English in fostering cooperation and competitiveness, but also warning against the risks of domain loss and the lack of English proficiency on the part of lecturers and students.

A close examination of how English is used in EMI programmes would greatly enhance our understanding of the challenges faced by the actors involved, providing empirical validation for perceptions as well as data to inform support materials. Nevertheless, while there have been a certain number of attitudinal studies, investigations of language use have been fewer and less homogenous methodologically, sometimes resorting to speech samples recorded in non-naturalistic settings (e.g. Airey 2015). As a result, although interesting insights have been provided, more results based on empirical data are still needed.

Corpora are invaluable tools for the systematic documentation of language use across disciplines and settings. Previous studies of stakeholder perceptions (Pulcini and Campagna 2015) and classroom observation (Drljača Margić and Vodopija-Krstanović 2017) have highlighted the difficulties sometimes experienced by lecturers in communicating intended meanings fluently and accurately. Drawing on these studies, in this contribution I will focus on the lexico-grammatical dimension of EMI lectures and investigate recurrent lexical bundles, i.e. sequences of words that display a statistical tendency to occur together.

I will compare lexical bundles retrieved from a small, specialised corpus of ten EMI lectures (c. 65,000 words) in the fields of the Physical Sciences and Engineering with lexical bundles retrieved from an equivalent sample in terms of disciplines and size taken from the *British Academic Spoken English* (BASE) corpus (Thompson and Nesi 2001).

The analysis of lexical bundles will provide information on their structure, functional properties and distribution within and across the corpora. This study is exploratory because the samples available do not allow for sweeping generalisations. Hence, the aim is to gain initial insights into how Italian EMI lecturers in scientific and technical fields use lexical bundles and how they compare with their colleagues teaching in settings where native English is the norm.

## 2. Lexical bundles in university lectures

Lexical bundles are strings of words of variable length (two or more items). These units have been labelled in different ways, such as

'clusters', 'lexical phrases' or 'routines', depending on whether the emphasis is on grammar, lexis or pragmatics. The term 'lexical bundle' is used here, following Biber *et al.* (1999), because the approach taken to identifying multi-word units is frequency-driven rather than based on perceptual salience (see Simpson 2004).

Lexical bundles tend not to be idiomatic in meaning, nor are they necessarily complete structural units; rather, they often function as a bridge between units, occurring at phrase or clause boundaries. For this reason, lexical bundles are indicators of fluent language use and represent a "key way of shaping text meanings" (Hyland 2012: 152).

Lexical bundles vary across registers (Biber *et al.* 1999; Biber, Conrad, Cortes 2004). Biber *et al.* (1999) show that in conversation, as opposed to academic prose, they are more diverse and constitute a greater proportion of discourse. In addition, while in conversation lexical bundles mainly convey stance (e.g. *I don't know what, I don't think so*) or function as interaction markers (e.g. *what do you think*), in academic texts they tend to perform referential (e.g. *at the end of the*) or discourse organising roles (e.g. *on the other hand*) (Biber *et al.* 2003).

Lexical bundles in academic speech have received considerable attention. A pioneering study was carried out by DeCarrico and Nattinger (1998) on the contribution of 'lexical phrases', i.e. strings of varying length such as *that goes without saying*, to the comprehension of academic lectures. Subsequently, thanks to the adoption of corpus-based approaches, lexical bundles were identified empirically, discovering recurring sequences that would otherwise have been difficult to identify. Various investigations have been conducted comparing university teaching, conversation and/or other academic registers (Biber, Conrad, Cortes 2004; Biber and Barbieri 2007). These studies demonstrate that university teaching employs lexical bundles to a greater extent than any other register. This type of spoken academic discourse is characterised by items found in both conversation and academic prose, because teaching combines the communicative needs of involved spoken interaction and the informational concerns of academic writing. Finally, classroom teaching relies more than other registers on bundles that organise discourse (Biber, Conrad, Cortes 2004: 397).

Studies on the in-text distribution of lexical bundles in lectures have also been conducted, focusing in particular on their discourse-

structuring role. Scholars have investigated the contribution of lexical bundles to cohesion (Nesi and Basturkmen 2006) and the relationship between lexical bundles and the macro-structure of discourse (Csomay 2013). Nesi and Basturkmen (2006) argue for the importance of making non-native students aware of discourse structuring lexical bundles in order to help them process long stretches of talk. Csomay (2013) shows that the distribution and function of lexical bundles shift in accordance with the macro-phases of lecture discourse: at the start of classes, lexical bundles tend to be used for interpersonal meanings and to refer to time, place or text; then when the more instructional phase of the lecture begins, there is a shift to referential and discourse organising bundles.

The evidence provided in previous research will be compared to the data obtained in the present study to shed some light on the use of lexical bundles by Italian EMI lecturers, paying particular attention to whether the distribution of forms and functions is in line with the overall expectations for the register of university teaching.

## 3. Materials and methods

This study is based on two small, specialised corpora of lectures delivered in English in Anglophone and non-Anglophone university settings (Table 1). Each corpus includes ten lectures for a total of around 65,000 words. The EMI classes are taken from a larger corpus currently being compiled and transcribed at the University of Torino, labelled *Turin EMI Lecture Corpus* (TEMILC). The transcriptions for this study were selected so as to constitute a sample that could be compared with lectures taken from the *British Academic Spoken English* (BASE) corpus[1], which contains 160 lectures and 40 seminars recorded at the Universities of Warwick and Reading. The sub-corpora obtained from TEMILC and BASE will be referred to as TEMILC-10 and BASE-10.

[1] https://warwick.ac.uk/fac/soc/al/research/collections/base/, last accessed June 29, 2018.

TABLE 1
Corpora for analysis

| *TEMILC-10*[a] | *Tokens* | *BASE-10*[b] | *Tokens* |
|---|---|---|---|
| LELUNDAI02 | 11,635 | pslct003 | 5,423 |
| LELUNDCA01 | 4,282 | pslct007 | 4,849 |
| LELUNDCH02 | 5,654 | pslct008 | 6,194 |
| LELUNDCS01 | 8,775 | pslct010 | 6,977 |
| LELUNDEC02 | 5,729 | pslct011 | 7,496 |
| LELUNDMA01 | 5,807 | pslct022 | 9,103 |
| LELUNDPH01 | 5,918 | pslct025 | 5,549 |
| LELGRADFL01 | 6,489 | pslct036 | 6,614 |
| LELGRADMD02 | 8,928 | pslct029 | 6,076 |
| LELGRADOM02 | 4,519 | pslct030 | 5,122 |
| Total | 67,736 | | 63,403 |

[a] TEMILC-10 includes seven undergraduate and three graduate lectures, each delivered by a different lecturer. The codes indicate the genre of the lecture and the size of its audience (LEL=Lecture/Large); the degree level (undergraduate or graduate); the field (e.g. CH=Chemistry); and the lecture number (e.g. 02=second lecture in the corpus in a given discipline). The following disciplines are included in TEMILC-10: Ambient Intelligence (AI); Computer Architecture (CA); Chemistry (CH); Computer Science (CS); Electronic Circuits (EC); Mathematical Analysis (MA); Physics (PH); Formal Languages and Compilers (FL); Mobile Application Development (MD); Optimisation Methods and Algorithms (OM).
[b] BASE-10 comprises 10 undergraduate lectures, each by a different lecturer, in the following disciplines (codes in parentheses indicate the lecture code in BASE): Chemistry (pslct003); Computer Science (pslct007, pslct008 and pslct010); Cybernetics (pslct011); Engineering (pslct022); Mathematics (pslct025); Mathematical Statistics (pslct036); Meteorology (pslct029 and pslct030).

Some considerations on the strengths and weaknesses of the design of the two comparable corpora are necessary. In a study of three-word strings in academic writing across a range of disciplines, Oakey (2009) found that the rank order of the ten most frequent strings in each discipline presented only minimal differences when obtained from comparable corpora with the same number of tokens but a different number of texts (i.e. isolexical) or from comparable corpora with the same number of texts but a different number of tokens (i.e. isotextual). This result indicates that the choice of one or the other corpus design to compare what the most frequent lexical bundles are in two or more situations of use will have little effect on the findings. Considering that the corpora used here are isotextual and almost isolexical, the risk of obtaining skewed data is minimal. Nevertheless, it should be pointed out that Oakey was dealing with much larger corpora than the ones used here and that some internal variation exists within TEMILC-10 and BASE-10. Hence, when

needed for comparative purposes, frequencies will be normalised to 100,000 words.

On the other hand, when it comes to analysing the distribution of the discourse functions of three-word strings (and likely longer strings, too), Oakey (2009) suggests that isotextual corpora should be used, because discourse functions are dependent on the overall organisation of the communicative event: in isolexical corpora, which are unbalanced in terms of the number of texts, there may simply be more occasions for the use of a specific function, a fact that will introduce a bias in the results. This consideration is particularly valid for formulaic communicative events, which present predictable rhetorical moves. However, an equal number of texts may not be a sufficient criterion to achieve comparability in highly variable genres such as the lecture, where it is difficult to ensure that the same amount of different 'discourse phases[2]' is present across different classes. Hence, in addition to choosing isotextual corpora, several contextual variables were controlled for in order to limit the effect of factors that may have an impact on the rhetorical organisation of lectures: both corpora only include monologic lectures; in both settings, lectures are delivered to large audiences (i.e. more than 40 students); all the lectures are mid-course classes, introductory and concluding sessions being characterised by distinct rhetorical features; finally, both corpora comprise lectures in the fields of the Physical Sciences and Engineering. One potentially impactful factor that could not be controlled for is whether lecturers rely on inductive or deductive reasoning, an aspect which may influence the type and function of lexical bundles. Hence, this point needs to be taken into account in the discussion of results.

As regards the procedure for the identification of lexical bundles, items were retrieved using the Word list function in Sketch Engine[3] (Kilgarriff, Rychly, Smrz, Tugwell 2004), which allows users to set specific parameters for an 'n-gram' search. In accordance with most empirical research on lexical bundles in academic spoken registers (e.g. Biber, Conrad, Cortes 2004; Nesi and Basturkmen 2006; Barber and Barbieri 2007; Csomay 2013), four-word sequences were considered. The cut-off point to identify lexical bundles is

[2] See Young (1994) for a model accounting for the macro-structure of university lectures.

[3] http://www.sketchengine.co.uk, last accessed June 29, 2018.

typically measured in normalised frequencies per million words (pmw). However, as has been observed (Biber, Conrad, Cortes 2004; Hyland 2012), the minimum threshold for bundle inclusion is a rather arbitrary parameter, which varies according to the size of the corpus (from 10 to over 100 occurrences pmw). In this study, the minimum raw frequency is six occurrences, which corresponds to a cut-off point of more than 80 occurrences pmw. This high normalised frequency is related to the small size of the corpora. The "inflated rate of occurrence" (Biber and Barbieri 2007: 268) obtained in small corpora has two consequences: on the one hand, the four-word strings found require corroboration in a corpus of one million words to fully qualify as 'lexical bundles'; on the other hand, four-word strings with low raw frequency (yet with high normalised rate) may appear repeatedly only in a few texts, thus reflecting the idiosyncrasies of some speakers. Concerning the first issue, in this paper it is only possible to regard the four-word strings identified with particular caution, bearing in mind that some of them may not turn out to be proper 'lexical bundles' in a one-million word corpus. As for the risk of including idiosyncratic uses, in the present study it was reduced by only considering the lexical bundles with a minimum raw frequency of six that appear in at least three lectures.

The decision to analyse bundles occurring in a minimum of three lectures was made following Biber and Barbieri (2007). In their study of lexical bundles in university spoken and written registers, they used different distributional thresholds according to the size of the corpus. For their 50,000-word corpora (i.e. office hours, 11 texts[4], and course management, 21 texts), they required bundles to occur in at least three different texts. Biber and Barbieri did not take the situation of use or the total number of texts into account in the identification of bundles; they set the parameter of 'at least three texts' uniquely based on corpus size. Distinct genres (e.g. highly dialogic office hours vs. monologic lectures) or corpora composed of dissimilar amounts of communicative events (11 office hours texts, 21 course management texts or the 10 lectures in this study) might require different criteria for the identification of lexical

[4] An office hour 'text' is an event that may comprise more than one meeting between a given faculty member and the students that go to see him/her during office hours (Biber and Barbieri 2007: 266).

bundles. Thus, the issue of what it takes for a four-word sequence to qualify as a lexical bundle in corpora smaller than one million words remains unresolved. For want of a better parameter, the criterion established by Biber and Barbieri (2007) of 'at least three texts' in a corpus of around 50,000 words was also adopted in the present study, which relies on samples totalling approximately 65,000 words.

## 4. Results

### 4.1. Frequency of lexical bundles

Table 2 shows that the lectures in BASE-10 contain 31 different types of four-word strings, while TEMILC-10 contains 26. Since the cut-offs for the two corpora were identical, this discrepancy indicates a slightly lower degree of variety in the EMI lectures examined. The total number of tokens (per hundred thousand words) differs, too, with the BASE-10 lecturers showing greater reliance on ready-made sequences. The gap between the two corpora is statistically significant[5]. However, it should be pointed out that the effect size measure[6] is very small (i.e. 0.00001), indicating that "something is really happening in the world, but [it can only be seen] through careful study" (Walker 2007, para. 6). This aspect needs to be taken into account in the interpretation of data.

TABLE 2
Number of lexical bundle types and tokens in the two corpora

| | *Types* | *Tokens* | *Tokens per 100,000 words* |
|---|---|---|---|
| TEMILC-10 | 26 | 219 | 323 |
| BASE-10 | 31 | 261 | 412 |

The results suggest that, overall, the lectures in BASE-10 are more

[5] Statistical analysis was conducted using Rayson's Log-likelihood (LL) calculator (http://ucrel.lancs.ac.uk/llwizard.html, last accessed June 29, 2018). For the raw tokens in Table 2, the LL value is 6.98, which is significant at $p<0.01$.

[6] The measure of effect size indicates how substantial the effect on performance is of the statistically significant differences observed. The effect size measure for Log-likelihood (ELL) was calculated using Rayson's ELL calculator (http://ucrel.lancs.ac.uk/llwizard.html, last accessed November 27, 2018).

formulaic than the EMI ones. If we consider that "control of [...] multi-word expressions" is a sign of "fluent linguistic production" (Hyland 2012: 150), the data presented here seem to provide initial empirical validation to the often reported reduced fluency of EMI teacher talk (Airey 2015; Pulcini and Campagna 2015; Drljača Margić and Vodopija-Krstanović 2017). On the other hand, the fact that TEMILC-10 has fewer types than BASE-10 may be related to the lower lexical variety sometimes observed in EMI lectures (Drljača Margić and Vodopija-Krstanović 2017).

## 4.2. Types

Table 3 shows the complete list of four-word strings in the two corpora.

TABLE 3
Four-word lexical bundles in the corpora: raw tokens, normalised frequencies (per 100,000 words), and number of lectures featuring the bundle

| *TEMILC-10* | | | | *BASE-10* | | | |
|---|---|---|---|---|---|---|---|
| *lexical bundle* | *raw* | *norm.* | *No. of lect.s* | *lexical bundle* | *raw* | *norm.* | *No. of lect.s* |
| the value of the | 14 | 20.67 | 4 | is going to be* | 27 | 42.58 | 6 |
| we are going to | 13 | 19.19 | 4 | and so on and | 14 | 22.08 | 5 |
| the end of the* | 13 | 19.19 | 4 | than or equal to* | 13 | 20.50 | 4 |
| in the case of* | 12 | 17.72 | 4 | if you want to* | 13 | 20.50 | 4 |
| we can say that | 10 | 14.76 | 4 | in terms of the* | 13 | 20.50 | 5 |
| if you have a* | 10 | 14.76 | 5 | or something like that* | 11 | 17.35 | 3 |
| at the end of* | 10 | 14.76 | 4 | going to be a* | 9 | 14.19 | 6 |
| in this case we | 9 | 13.29 | 7 | so we have to | 8 | 12.62 | 3 |
| i would like to | 9 | 13.29 | 4 | the same as the | 8 | 12.62 | 5 |
| and this is the* | 9 | 13.29 | 5 | are we going to* | 8 | 12.62 | 3 |
| you can have a | 8 | 11.81 | 3 | so this is a | 8 | 12.62 | 4 |

(*continued on next page*)

TABLE 3 (*continued from previous page*)
Four-word lexical bundles in the corpora: raw tokens, normalised frequencies (per 100,000 words), and number of lectures featuring the bundle

| *TEMILC-10* | | | | *BASE-10* | | | |
|---|---|---|---|---|---|---|---|
| so this is a | 8 | 11.81 | 3 | and things like that* | 8 | 12.62 | 4 |
| so that you can | 8 | 11.81 | 3 | a little bit more* | 8 | 12.62 | 6 |
| if you want to* | 8 | 11.81 | 4 | the end of the* | 7 | 11.04 | 3 |
| have a lot of* | 8 | 11.81 | 3 | to be able to* | 7 | 11.04 | 4 |
| we are talking about | 7 | 10.33 | 4 | less than or equal | 7 | 11.04 | 4 |
| we are able to | 7 | 10.33 | 3 | if we have a | 7 | 11.04 | 5 |
| this is an example | 7 | 10.33 | 5 | a little bit of* | 7 | 11.04 | 5 |
| the direction of the | 7 | 10.33 | 3 | we have to be | 6 | 9.46 | 5 |
| with respect to the | 6 | 8.86 | 3 | i was talking about | 6 | 9.46 | 3 |
| the ratio between the | 6 | 8.86 | 3 | in the middle and | 6 | 9.46 | 3 |
| so this is the | 6 | 8.86 | 5 | the other way round | 6 | 9.46 | 3 |
| is exactly the same | 6 | 8.86 | 4 | is the sum of | 6 | 9.46 | 4 |
| is an example of | 6 | 8.86 | 3 | so this is the | 6 | 9.46 | 3 |
| and this is a* | 6 | 8.86 | 3 | if you have a* | 6 | 9.46 | 5 |
| and in that case | 6 | 8.86 | 3 | have a look at | 6 | 9.46 | 3 |
| | | | | going to talk about | 6 | 9.46 | 3 |
| | | | | and the reason for | 6 | 9.46 | 4 |
| | | | | and so on so | 6 | 9.46 | 4 |
| | | | | and in this case | 6 | 9.46 | 3 |
| | | | | a bit of a | 6 | 9.46 | 4 |

Of the 31 types in BASE-10, 13 items (41%, marked with an asterisk in Table 3) are listed in identical form as frequent lexical bundles in university classroom teaching by Biber, Conrad, Cortes (2004: 384ff). Since these scholars used a corpus of more than one million words, those 13 items may be safely regarded as fully-fledged lexical bundles. By contrast, out of the 26 types in TEMILC-10, 8 (30%, marked with an asterisk, too) are present in Biber, Conrad, Cortes's (2004) catalogue, indicating that a greater amount of four-word strings in TEMILC-10 needs corroboration in larger corpora.

In both TEMILC-10 and BASE-10, most of the bundles that do not feature in the classroom teaching list are either slightly different from those identified by Biber, Conrad, Cortes (2004) (e.g. *if we have a* instead of *if you have a*) or influenced by the broad disciplinary field, i.e. the Physical Sciences and Engineering. The impact of this aspect is observable in some lexical choices (e.g. *the ratio between the* in TEMILC-10; *is the sum of* in BASE-10) and arguably in the strings that contribute to argumentative or explanatory routines (e.g. *in this case we* in TEMILC-10 and *and in this case* in BASE-10).

Looking at the distribution of types across the lectures, no major differences between the two corpora were noticed. In both cases, most of the instances occur in less than half the lectures and there is no bundle that appears across the board. In BASE-10, the bundles that occur in more than half the lectures are also among the most frequent lexical bundles in the register of university teaching (i.e. *is going to be; going to be a; a little bit more*). On the other hand, in TEMILC-10, the only type appearing in more than five classes, *in this case we*, does not feature among the common lexical bundles of the university teaching register. It is a device used by lecturers to focus the attention of students on the characteristics of the issue at hand in comparison to some other issue (e.g. *we still need a multiplexer but* ***in this case we*** *have three different possibilities* [LELUNDCA01]). The data available are too limited to draw any conclusions on these patterns, but the hypothesis may be formulated that *in this case we* is a string functional to the 'top-down logic' of deductive reasoning, which may reflect a particular pedagogical culture.

### 4.3. Structural patterns

All the items retrieved were grouped structurally into three broad

categories, following Bychkovska and Lee (2017). The first group (VP-based) is composed of clausal structures, i.e. those that incorporate verb phrases (e.g. *we are going to*) and dependent clause fragments (e.g. *so that you can*). The second set (NP-based) comprises noun phrases, including those with post-modifier fragments (e.g. *the end of the*. The third class (PP-based) is composed of prepositional phrases (*in the case of*). The 'Other' category contains strings that do not fit perfectly into any of the other groups, such as adjectival and adverbial expressions (e.g. *less than or equal* and *and so on and*). Table 4 shows the distribution of the main structural categories.

TABLE 4
Distribution of structural categories in the corpora

| | *Types (raw)* | | *Tokens (raw)* | | |
|---|---|---|---|---|---|
| | *TEMILC-10* | *BASE-10* | *TEMILC-10* | *BASE-10* | *LL*[a] |
| VP-based | 17 | 15 | 136 | 129 | 0.01 |
| NP-based | 4 | 9 | 40 | 67 | 8.79* |
| PP-based | 5 | 3 | 43 | 25 | 3.71 |
| Other | 0 | 4 | 0 | 40 | 58.14** |

[a] LL is used to compare the frequencies of tokens across two corpora, not the number of types. For this reason, the LL values in Tables 4 and 6 refer to tokens only.
* significant at p<0.01, ELL: 0.00002; ** significant at p<0.0001, ELL: 0.00015.

In both corpora, VP-based constructions are the most frequent class; nevertheless, a sizable proportion of occurrences are phrasal in nature, being composed of NP/PP-based strings and other clusters. Overall, this distribution reflects the characteristics of the register of university teaching, which draws on both conversation, where a predominance of VP-based and dependent clause bundles is found, and academic prose, where most bundles are NP/PP-based (Biber *et al.* 1999; Biber, Conrad, Cortes 2004).

Looking at the structural categories contrastively across corpora, it can be noticed that the EMI lecturers resort to NP-based and 'other' strings to a significantly lower extent than their colleagues in BASE-10. This mainly depends on the absence of expressions that are typically associated with the register of classroom teaching (Biber, Conrad, Cortes 2004: 387), such as the vagueness markers *or something like*

*that*, *and things like that*, *and so on and (so forth)* and hedges such as *a bit of a* and *a little bit of*. Since both vagueness markers and hedges perform interpersonal functions that promote pragmatically efficient communication, their absence may partly explain the impression of reduced "expressiveness" and "level of spontaneity" (Drljača Margić and Vodopija-Krstanović 2017: 90) sometimes observed in EMI.

### 4.4. Discourse functions

In order to analyse the discourse functions of lexical bundles, the taxonomy developed by Biber, Conrad, Cortes (2004) was adopted. These authors distinguish three primary functions, each of which subsumes more specific uses: stance bundles, discourse organisers and referential expressions. Stance bundles convey epistemic evaluations, attitudes or modal meanings. Discourse organisers make explicit reference to previous or upcoming stretches of discourse. Referential expressions specify attributes of various kinds pertaining to the physical or abstract entities being talked about.

In this study, the functions of lexical bundles were identified by exploring the meaning conveyed by occurrences in context. In most cases, the form-function associations identified by Biber, Conrad, Cortes (2004) were confirmed. Multifunctional strings were categorised according to their primary or most common function.

Table 6 shows that the proportional distribution of functions in the two corpora is approximately the same. The most frequent function is Referential (72% in TEMILC-10; 63% in BASE-10), followed by Stance Taking (25% in TEMILC-10; 27% in BASE-10) and Discourse Organising (3% in TEMILC-10; 10% in BASE-10). This distribution shows that the lectures in TEMILC-10 and BASE-10 are informationally dense and that they prioritise sequences normally associated with academic prose. This trait might depend on the disciplinary field. In his study of lexical bundles and disciplinary variation in academic writing, Hyland (2012: 164) found that in science and engineering texts there was a much higher concentration of research-oriented (i.e. ideational) bundles. He explains this feature as a result of the greater focus of these disciplines on "the description or specification of research objects or context" (p. 164). This interpretation may also apply to the university lectures under scrutiny.

TABLE 5
Distribution of functional categories and subcategories

| *Categories* | *Sub-categories* | *Types* | | *Tokens (per 100,000 words)* | | |
|---|---|---|---|---|---|---|
| | | *TEMILC -10* | *BASE -10* | *TEMILC -10* | *BASE -10* | *LL* |
| Stance bundles[a] | Epistemic stance | 0 | 0 | 0 | 0 | - |
| | Attitudinal/ modality | 6 | 6 | 55 (81.20) | 70 (110.40) | 2.93 |
| | Total | 6 | 6 | 55 (81.20) | 70 (110.40) | 2.93 |
| Discourse organisers[b] | Topic introduction/ focus | 0 | 3 | 0 | 20 (31.54) | 29.07*** |
| | Topic elaboration/ clarification | 1 | 1 | 7 (10.33) | 6 (9.46) | 0.03 |
| | Total | 1 | 4 | 7 (10.33) | 26 (41.01) | 12.93** |
| Referential[c] | Identification/ focus | 6 | 2 | 42 (62.01) | 14 (22.08) | 12.86** |
| | Imprecision | 0 | 4 | 0 | 39 (61.51) | 56.69*** |
| | Attributes | 11 | 13 | 92 (135.82) | 99 (156.14) | 0.93 |
| | Time/place/ text reference | 2 | 2 | 23 (33.96) | 13 (20.50) | 2.19 |
| | Total | 19 | 21 | 157 (231.78) | 165 (260.24) | 1.08 |
| **Total** | | 26 | 31 | 219 (323.31) | 261 (411.65) | 6.98* |

[a] TEMILC-10: *i would like to* (desire); *we are going to* (intention/prediction); *we can say that, so that you can, you can have a, we are able to* (ability). BASE-10: *if you want to* (desire)*; so we have to, we have to be* (obligation/directive/necessity); *is going to be*, *going to be a* (intention/prediction); *to be able to* (ability).
[b] TEMILC-10: *we are talking about* (topic elaboration/clarification). BASE-10: *are we going to, going to talk about, have a look at* (topic introduction/focus); *i was talking about* (topic elaboration/clarification).
[c] TEMILC-10: *and this is the, so this is a, this is an example, and this is a, so this is the, is an example of* (identification/focus); *the value of the, in the case of, if you have a, in this case we, have a lot of, if you want to, the direction of the, the ratio between the, and in that case, is exactly the same, with respect to the* (specification of attributes); *the end of the, at the end of* (time/place/text reference). BASE-10: *so this is a, so this is the* (identification/focus); *and so on and, and so on so, or something like that, and things like that* (imprecision); *in terms of the, than or equal to, the same as the, a little bit more, a little bit of, less than or equal, if we have a, if you have a, a bit of a, is the sum of, the other way round, and in this case, and the reason for* (specification of attributes); *the end of the, in the middle and* (time/place/text reference).
* significant at p<0.01, ELL: 0.00001; ** significant at p<0.001, ELL: 0.00004 and 0.00003 (in the order given in the table); *** significant at p<0.0001, ELL: 0.00015.

A comparison of TEMILC-10 and BASE-10 reveals statistically significant differences between the two corpora. First, the lecturers in TEMILC-10 use considerably fewer discourse organising bundles. Second, although referential expressions do not differ in frequency as a whole, identification/focus and imprecision bundles are significantly underused by the TEMILC-10 lecturers. The remainder of this section will analyse the uses observed in the two corpora in context, trying to elucidate possible reasons for these distributional patterns.

Stance comprises two main categories: epistemic stance and attitudinal/modal stance. Neither corpus contains four-word epistemic stance bundles. However, the three-word bundle *I don't know*[7] is very frequent in both corpora. Probably, the small size of the samples did not allow this string to combine with other recurring items forming a sufficient number of four-word units. Hence, a replication of this study in larger corpora might provide a different picture regarding epistemic stance bundles. On the other hand, attitudinal/modality sequences are rather frequent. In both corpora, they convey desire, intention/prediction and ability, and are mainly related to discussions around the teaching subjects (examples 1 and 2). In TEMILC-10, however, intention/prediction bundles co-occur with expressions announcing discourse goals (example 3). This use was also noted in Biber, Conrad, Cortes (2004: 390) and Biber and Barbieri (2007: 275).

1) the vibration **is going to be** along this way [pslct022]

2) the reason why these flowchart are good well it's because **we are able to** translate this into a eh directly into a c programme [LELUNDCS01]

3) […] what **we are going to** discuss in the next eh part of the morning […] [LELUNDEC02]

In BASE-10 only, attitudinal/modality bundles also convey necessity (example 4). The expression of necessity rather than obligation or directives, as noted instead in classroom teaching (Biber, Conrad, Cortes 2004: 390; Biber and Barbieri 2007: 275), may be due to the

[7] The string *I don't know* might be considered as a four-word bundle if the contracted verb form were counted as two words. For comparative purposes with Biber, Conrad, Cortes (2004), who treat contracted verb forms as a single unit, *I don't know* is considered a three-word bundle in the present study.

predominantly monologic and theoretical nature of the lectures collected: rather than asking students to accomplish certain tasks, instructors interact with them by engaging them in the unfolding discourse.

4) **so we have to** be much stricter about what sets are allowed in the scheme of things [pslct025]

Discourse organising bundles are divided into two sub-groups. The first includes items that overtly signal the introduction of topics or the shift to a new topic (example 5). The second is composed of elements that elaborate on a given theme or provide clarifications. Only the second group was identified in both TEMILC-10 and BASE-10 (examples 6 and 7), while no instances of lexical bundles explicitly introducing topics were found in TEMILC-10. This is the reason why discourse organising bundles are less frequent overall in the EMI lectures examined.

5) so let's **have a look at** how that works [pslct010]

6) you could simplify one plus beta 1 and one plus beta 2 because **we are talking about** an integrated eh solution [LELUNDEC02]

7) that's one thing i didn't make clear was that **i was talking about** what i call simply connected regions [pslct025]

Referential bundles include four more specific categories: identification/focus, imprecision, specification of attributes and time/place/text references. No differences between the two corpora were noticed in terms of the frequency of occurrence of bundles specifying attributes and expressing time/place/text references. In addition, not only are many of them (almost) identical in form (e.g. *and in this case* or *and in that case; the end of the*), but the specific meanings in context are also shared. For example, in the case of attribute specifications, both the TEMILC-10 and the BASE-10 lecturers use bundles to refer to amounts (e.g. *have a lot of; less than or equal*), to frame tangible attributes (e.g. *the direction of the; is the sum of*) and to signal logical relationships (e.g. *with respect to the; in terms of the*).

On the other hand, EMI lecturers use significantly more items that identify the entities being talked about (example 8). The reasons

why identification/focus strings are so numerous in TEMILC-10 are related to the higher number of types and their co-occurrence with metadiscourse (example 9), a feature not attested in BASE-10 but observed by Biber, Conrad, Cortes (2004: 394).

8) **so this is a** thermal machine which has a cycle produces work okay? [LELUNDPH01]

9) **and this is a** quite important point okay? [LELUNDMA01]

Another significant difference between the two corpora is the absence of imprecision bundles in TEMILC-10, as noted in Section 4.3. Although Biber, Conrad, Cortes (2004: 387) report that imprecision bundles are extensively used in the register of university classroom teaching in English, the EMI lecturers in the corpus may regard these expressions as too informal to be employed in university lectures, because they may be transferring their expectations based on the more formal interaction typical of Italian-medium instruction.

## 5. Concluding remarks

The goal of the present study was to provide initial insights into the use of lexical bundles in EMI lectures. For this purpose, four-word strings were retrieved from two small, specialised corpora of lectures in technical and scientific fields delivered in Anglophone (the UK) and non-Anglophone (Italy) university settings. One of the reasons why lexical bundles are worth exploring is their contribution to fluency, an aspect which may be a problem in EMI and which is assessed in existing tests to measure lecturers' ability to teach in English (e.g. the Test of Oral English Proficiency for Academic Staff, TOEPASS; see Dimova 2017).

The findings of this study show that the EMI lectures examined conform to the expectations for the university teaching register in terms of the distribution of structural categories, with a predominance of VP-based bundles and a similar amount of NP/PP-based ones taken together. In addition, the TEMILC-10 lectures are very similar to the BASE-10 ones in the markedly higher use of referential strings over stance and discourse organising items, a

feature that may be attributed to the impact of the broad disciplinary areas of the Physical Sciences and Engineering.

Despite these similarities, some statistically significant differences were noted between TEMILC-10 and BASE-10. The EMI lectures in TEMILC-10 are less formulaic than the lectures in BASE-10 in terms of both the number of types identified and their frequency of occurrence. In addition, there are more 'unconventional' four-word strings in TEMILC-10 which were not reported in previous studies of lexical bundles in academic lectures. Finally, the analysis of structural types and discourse functions revealed that some items typically associated with the university teaching register are absent in the EMI lectures analysed, namely imprecision bundles (*or something like that*), hedges (*a little bit of*) and topic introduction bundles (*have a look at*).

The items and distributional patterns identified in this study need corroboration in larger corpora. Further investigations should confirm whether all the four-word strings retrieved qualify as lexical bundles proper and whether the distributional patterns of TEMILC-10 can be generalised. In addition, while TEMILC-10 and BASE-10 differ in the frequency of four-word strings, the difference is not large enough to be seen 'with the naked eye' (Walker 2007, para. 6), as revealed by the small effect size measure obtained for the overall tokens. Therefore, another point in need of verification is whether the lower use of lexical bundles in EMI is a feature that actually makes a substantial difference to performance.

The issue of corpus size is related to the challenging and time-consuming tasks of obtaining and transcribing spoken data. One solution could be to 'join forces' among researchers and make materials available to other colleagues, although this option might be problematic due to privacy issues. It is hoped, at least, that in the future more studies on language use in EMI will be conducted based on more principled collections of texts than has been the case so far. Research in this field should establish shared guidelines for data collection and make sure that contextual variables are recorded to allow comparisons. In this way it will be possible to document language use in EMI across settings and disciplines in a more rigorous manner.

Despite the objective limits of any exploratory analysis, some of the results obtained here stimulate reflections on what the

applications of studies on lexical bundles may be in teacher training and in pre-sessional courses preparing students for English-medium education. A general reason for including lexical bundles in teaching is offered by Biber and Barbieri (2007: 284), who argue that 'marked' language forms (often highly salient but not necessarily recurrent) are acquired more easily than forms that are not perceptually salient. Hence, lexical bundles, which are "not always obvious to the listener or the speaker" (Neely and Cortes 2009: 21), are eligible constructions for overt instruction.

Considering that in TEMILC-10 topic introduction bundles were not found at all, teacher trainers may devise activities that raise awareness among lecturers of the need to assist students with the explicit marking of topic introductions. Lecturers might be asked to conduct self-reflective activities by filling in *ad hoc* grids at the end of their lectures on how they behaved in relation to the introduction of topics and the transition from one topic to the next. Subsequently, using authentic materials, such as corpus transcription excerpts, lecturers could be asked to identify an inventory of forms for the introduction of topics. Teacher trainers might point out that while impersonal discourse markers such as *so* or *now* may be used, more explicit forms are more effective in enhancing lecture comprehension (Neely and Cortes 2009: 22-23). Lecturers might then be led to notice that the more syntactically elaborate forms for topic introductions often contain recurrent lexical bundles. Finally, lecturers could be exposed to a variety of lexical bundle forms for topic introductions and helped to consolidate use of these forms through cloze-type activities, taken from concordance lines, requiring the insertion of the most appropriate bundle (Neely and Cortes 2009: 33).

A further finding of this study is the absence in EMI lectures of bundles conveying pragmatic meanings, particularly vagueness markers and hedges. In this case, lexical bundles could be exploited by teacher trainers not so much in the sense of explicitly encouraging their use, but rather for the purpose of stimulating discussion of the beliefs the lecturers have regarding degrees of formality in lectures and whether they consider their current practices in EMI to be successful.

The activities that can be conducted on lexical bundles with lecturers could also be proposed, with due adjustments, to students

during pre-sessional EAP courses. However, it is essential that teacher trainers and EAP practitioners collaborate on the design of syllabi identifying converging goals, so as to better contribute to EMI quality assurance and support.

## Acknowledgments

The author would like to thank the anonymous reviewers for their insightful suggestions.

## References

Airey, John, 2015, "From Stimulated Recall to Disciplinary Literacy: Summarizing Ten Years of Research into Teaching and Learning in English", in S. Dimova, A.K. Hultgren, C. Jensen (eds), *English-Medium Instruction in European Higher Education*, De Gruyter, Berlin, pp. 157-176.

Biber, Douglas and Barbieri, Federica, 2007, "Lexical Bundles in University Spoken and Written Registers", *English for Specific Purposes* 26 (3), pp. 263-286.

Biber, Douglas, Conrad, Susan, Cortes, Viviana, 2003, "Lexical Bundles in Speech and Writing: An Initial Taxonomy", in T. McEnery, P. Rayson, A. Wilson (eds), *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*, Peter Lang, Frankfurt am Main, pp. 71-93.

Biber, Douglas, Conrad, Susan, Cortes, Viviana, 2004, "*If You Look at...*: Lexical Bundles in University Teaching and Textbooks", *Applied Linguistics* 25 (3), pp. 371-405.

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

Broggini, susanna and Costa, francesca, 2017, "A Survey of English-Medium Instruction in Italian Higher Education. An Updated Perspective from 2012 to 2015", *Journal of Immersion and Content-Based Language Education* 5 (2), pp. 238-64.

Bychkovska, Tetyana and Joseph J., Lee, 2017, "*At the Same Time*: Lexical Bundles in L1 and L2 University Student Argumentative Writing", *Journal of English for Academic Purposes* 30, pp. 38-52.

Campagna, Sandra and Pulcini, Virginia, 2014, "English as a Medium of Instruction in Italian Universities: Linguistic Policies, Pedagogical Implications", in M.G. Guido and B. Seidlhofer (eds), *Textus. English Studies in Italy. Perspectives on English as a Lingua Franca*, 27 (1), pp. 173-90.

CORTES, VIVIANA, 2008, "A Comparative Analysis of Lexical Bundles in Academic History Writing in English and Spanish". *Corpora* 3(1), pp. 43-57.

CSOMAY, ENIKO, 2013, "Lexical Bundles in Discourse Structure: A Corpus-Based Study of Classroom Discourse", *Applied Linguistics* 34 (3), pp. 369-388.

DECARRICO, JEANETTE and NATTINGER R., JAMES, 1988, "Lexical Phrases for the Comprehension of Academic Lectures", *English for Specific Purposes* 7 (2), pp. 91-102.

DIMOVA, SLOBODANKA, 2017, "Life after Oral English Certification: The Consequences of the Test of Oral English Proficiency for Academic Staff for EMI Lecturers", *English for Specific Purposes* 46, pp. 45-58.

DRLJAČA MARGIĆ, BRANKA and VODOPIJA-KRSTANOVIĆ, IRENA, 2017, *Uncovering English-Medium Instruction: Glocal Issues in Higher Education*, Peter Lang, Bern.

HYLAND, KEN, 2012, "Bundles in Academic Discourse", *Annual Review of Applied Linguistics* 32, pp. 150-169.

KILGARRIFF, ADAM, RYCHLY, PAVEL, SMRZ, PAVEL, TUGWELL, DAVID, 2004, "The Sketch Engine", in G. Williams and S. Vessier (eds), *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6-10, 2004*, Université de Bretagne-Sud, Faculté des lettres et des sciences humaines, Lorient, pp. 105-116.

MOLINO, ALESSANDRA, 2017, "Repetition and Rephrasing in Physical Sciences and Engineering English-Medium Lectures in Italy", in C. Boggio and A. Molino (eds), *English in Italy: Linguistic, Educational and Professional Challenges*, Franco Angeli, Milano, pp. 182-202.

MOLINO, ALESSANDRA and CAMPAGNA, SANDRA, 2014, "English-Mediated Instruction in Italian Universities: Conflicting Views", *Sociolinguistica, Internationales Jahrbuch für europäische Soziolinguistik* 28, pp. 156-171.

NEELY, ELIZABETH and CORTES, VIVIANA, 2009, "*A little bit about*: Analyzing and Teaching Lexical Bundles in Academic Lectures", *Language Value*, 1 (1), pp. 17-38.

NESI, HILARY and BASTURKMEN, HELEN, 2006, "Lexical Bundles and Discourse Signaling in Academic Lecturers", *International Journal of Corpus Linguistics*, 11 (3), pp. 283-304.

OAKEY, DAVID, 2009, "Fixed Collocational Patterns in Isolexical and Isotextual Versions of a Corpus", in P. Baker (ed), *Contemporary Corpus Linguistics*, Continuum, London, pp. 140-58.

PULCINI, VIRGINIA and CAMPAGNA, SANDRA, 2015, "Internationalisation and the EMI Controversy in Italian Higher Education", in S. Dimova, A.K. Hultgren, C. Jensen (eds), *English-Medium Instruction in European Higher Education*, De Gruyter, Berlin, pp. 65-87.

SIMPSON, RITA, 2004, "Stylistic Features of Academic Speech: The Role of Formulaic Expressions", in T. Upton and U. Connor (eds), *Discourse in the Professions: Perspectives from Corpus Linguistics*, John Benjamins, Amsterdam, pp. 37-64.

THOMPSON, PAUL and NESI, HILARY, 2001, "The British Academic Spoken English (BASE) Corpus Project", *Language Teaching Research* 5 (3), pp. 263-264.

WALKER, IAN, 2007, "Null Hypothesis Testing and Effect Sizes", in *Statistics for Psychology*, available at http://staff.bath.ac.uk/pssiw/stats2/page2/page14/page14.html, last accessed November 27, 2018.

YOUNG, LYNNE, 1994 "University Lectures: Macro-structure and Micro-features", in J. Flowerdew (ed.), *Academic Listening: Research Perspectives*, CUP, Cambridge, pp. 159-76.