

Identifying Lexical Bundles in Secondary School Textbooks

Carola Ribeck

Språkbanken, Department of Swedish
University of Gothenburg, Sweden
carola.ribeck@gu.se

Abstract

The present paper describes the process of identifying *lexical bundles*, i.e. frequently recurring word sequences such as *by means of* and *in the end of*, in secondary school textbooks of history and physics. In its determination of finding *genuine lexical bundles*, i.e. the word boundaries between lexical bundles and surrounding arbitrary words, it proposes a new approach to come to terms with the problem of extracting overlapping bundles of different lengths. The results show that surprisingly few bundles are common to both subjects. The structural distribution across the subjects indicates that history uses more NP/PP-based and less dependent-clause-based bundles than physics. The comparative analysis manages to restrict this difference to the referential function. History almost only refers to phrases, i.e. within clauses, while physics much more tends to make references across clauses.

Keywords: lexical bundles, formulaic sequences, multi-word units, textbooks, secondary school

1. Introduction

The paper describes lexical bundles in secondary school textbooks of history and physics. It is the first report with a subject-contrastive perspective within this genre. It is also the first account of lexical bundles in a Swedish corpus. It should be regarded as an explorative pilot study aiming to find out to what extent lexical bundles are to be found in the texts the students have to read and if there are any structural or functional differences in the way they are used in the natural and social sciences.

The incentive for conducting the study is the awareness of the great importance language has for learning. Our school needs to offer support to all students at all levels, high-proficiency as well as low-proficiency, native as well as non-native. In order to achieve this, we all need a better understanding of the linguistic challenges of schooling.

1.1 Language and schooling

The language of schooling can be a difficult code for students to break, especially for second language learners. Cummins (1980) and Macken-Horarick (1996) go as far as describing school language as a completely different language from the language used in everyday communication. “*Students may be fluent in spoken English but still lack basic resources for reading and writing academic registers*” (Cummins 1980).

According to functional linguistic theory, learning a subject means learning its language. “*It is through the language of schooling that the concepts are construed, and through the development of control of academic registers that students develop consciousness about the concepts of advanced literacy*” (Schleppegrell 2004). Halliday and Martin (1993) describe the language of both natural and social sciences as characterised by lots of nominalisations; yet history is said to reason within sentences and natural science between clauses.

Gardner (2007) stresses the importance of teaching language with the most correct notion of what constitutes a *word*, i.e. the largest possible entities of language correlating to a specific meaning.

1.2 Multi-word units

The interest for recurrent word sequences stretches back to researchers like Palmer and Firth and the mid 1900’s. They used the notion of *collocations*, which Halliday some years later defined more precisely, highlighting the propensity of a lexical item to co-occur with one or more other words.

This sort of recurring word patterns have been studied under a number of other names throughout the years e.g. *conventionalized language forms*, *speech formulas*, *ready-made expressions*, *fixed expressions*, *pre-fabricated language*, *multi-word units/expressions* and *lexical bundles* (Cortes 2004, Strunkyte and Jurkunaite 2008). Multi-word units of various sorts play fundamental parts in natural languages. Acquiring an understanding for them is crucial for succeeding in any specialised linguistic community, e.g. academic education (Atkins and Rundell 2008, Gardner 2007).

Many recent studies have focused on the use of different kinds of multi-word units in the language of native vs. non-native speakers. Chen and Baker (2010) reports more NP-based bundles in native academic writing, while non-native writing has a tendency to over-generalise and favour “*certain idiomatic expressions and connectors*”. Nekrasova (2009) comes across less lexical bundles in the speech of lower-proficiency English learners than in the speech of native students and higher-proficiency learners.

In the quantitative corpus studies so far, one can distinguish two different procedures. The first draws on particular sequences of interest, which have been picked out before the study, because of their familiarity to native speakers or their high frequency. The second is based on the results of a program that can find sequences of some lengths above various cut-off frequencies.

1.3 Lexical bundles

Biber and Conrad (1999) was the starting point for large-scale corpus linguistic studies of *lexical bundles*, which here were defined as “*sequences of three or more words that show a statistical tendency to co-occur*”. Biber and his colleagues have over the years developed a method for identifying, characterising and analysing lexical bundles.

Lexical bundles can be seen as an extended conception of collocations, which usually refer to two-word sequences. However they differ from other kinds of multi-word units in three major aspects: first, they are extremely frequent, second, they have no idiomatic meaning and last, they are not perceptually striking in themselves (Biber and Barbieri, 2007). For a long time, lexical bundles escaped observation, as they often transcend structural boundaries.

Hyland (2007) upholds that since lexical bundles are transparent in meaning, they are crucial for construing coherence in a discourse. Strunkyte and Jurkunaite (2008) add that comprehension of lexical bundles improves the receptive process of reading.

Biber et al. (2004) look at university textbooks as one register, where they locate relatively few bundles, compared to, e.g. classroom teaching. The textbook bundles found are mostly NP/PP-based ones which are used for referential purposes. Biber and Barbieri (2007) show that there are characteristic sets of lexical bundles with specific discursive functions connected to different registers.

Strunkyte and Jurkunaite (2008) compare bundles in research articles in humanities and natural science. The language in humanities is found to be more structurally varied, while the language of natural science contains more text-organising bundles.

1.4 Genuineness

A returning question regarding lexical bundles is whether they are pre-fabricated (Wray 2002), as in forming wholes in our mental lexicons, regarding storage as well as retrieval. Biber and Barbieri (2007) hypothesise that “*high frequency is a reflection of pre-fabricated status*”.

This approach, of solely taking frequency into account, has been a target of criticism, inter alia by McEnery et al. (2006). Nekrasova (2009) maintains that high-frequent sequences are of different strengths: “*A bundle should be described in terms of its place on a continuum from more holistic to more compositional units.*”

There are few attempts to discriminate between, let us say, *genuine lexical bundles* from *non-genuine ones*¹. A genuine lexical bundle would be one that both structurally and functionally forms one piece of language (cp. Gardner 2007 above). To be of true linguistic interest, studies on lexical bundles have to aim for genuine targets.

2. Data and method

For this study, two sub-corpora from the Swedish corpus for textbooks, *OrdiL* (Lindberg and Johansson Kokkinakis 2007), were used: one with texts in physics and one with texts in history. Both corpora consisted of two textbooks from secondary school.

The physics books contained (54 334 + 77 318) = 131 652 words and the history books (45 215 + 30 524) = 75 739 words. So that enable comparisons, the absolute frequencies of bundles were normalised to reflect the number of occurrences per 300 000 words.

To identify the bundles, a search for n-grams of > 2 words that were found ≥ 2 times in both of the subject's

textbooks² was made by means of AntConc (<http://www.antlab.sci.waseda.ac.jp/>) and an additional script.

2.1 Overlaps

Since all bundles containing more than three words inevitably enclose bundles of shorter word lengths, the initial bundle lists were full of more or less overlapping sequences. Cortes (2004) explicitly gives this as the reason only to look at 4-word sequences, while Hyland (2008) simply leaves all overlaps.

In the spirit and determination of sorting out the most genuine lexical bundles, the present paper presents a new approach to come to terms with the problem of overlapping sequences. The method is semi-automatic-semi-manual. With the intention of removing all bundles but one, a program identified all overlapping groups and then proposed the one bundle to retain, according to the following order of precedence: *length > frequency*³.

In the subsequent manual step, it was decided whether to approve of the proposed bundle or choosing another one. At this point, an additional aspect was considered, namely *intuitive genuineness*.

2.2 Splits

For bundle groups of 3-5 words, strictly one bundle was kept. However exceptions were sometimes made for groups containing sequences of more than 5 words. If the longer sequence consisted of significantly more frequent non-overlapping shorter bundles, the longer sequence would be split, e.g.

```
32 vad är det (what is it)
6 vad är det som gör att
  (what is that makes)
20 som gör att (that makes)
```

2.3 Classification

This study uses Biber et al.'s (2004) structural and functional categories for classification. Structurally the bundles are placed into three catchall groups: *NP/PP-based (P-f)*, *VP-based (VP-f)* and *dependent clause based fragments (DC-f)*. The functional all-embracing labels are: *stance expressions*, *referential expressions* and *discourse organizers*.

The structural classification was made manually. The only adjustment to Biber et al. (2004) concerned a subset of “*noun phrase with other post-modifier fragment*”. To be more precise, bundles containing NP-fragments and relative subordinate clause fragments were labelled DC-f:s, rather than P-f:s. This was done in order to spot phrasal and clausal boundaries to the utmost possible extent.

The functional analysis that required studying the concordances could only be carried out for the 30 highest ranked bundles of every subject. A functional category, called *lexically based expressions*, had to be introduced to contain the bundles that had no other function than to express the lexical contents of the constituent words.

¹ Attia et al. (2010), Simpson-Vlach and Ellis (2010) and Zhang et al. (2006) being some exceptions to the rule.

² Demanding the sequences to be found across texts prevents mere idiosyncracies from being counted.

³ A longer sequence was suggested if the frequency was higher than 60% than that of a one word shorter sequence.

3. Result

The results are shown in this section and discussed in the next one. 21 bundles were common to both subjects.

3.1 Bundle density

The *bundle density* (BD) of the corpus was defined as the number of occurring bundles divided by the number of words. The higher the bundle density, the more likely it is to find a lexical bundle in a text of a certain word length. The *bundle variation* (BV) is the bundle density divided by the number of bundle types found. This measure is higher the more repetitive the text is regarding bundles.

	words	bundles (types)	bundles (occurrences)	bundle density	bundle variation
History	75739	124	3388	0,045	$3,6 \cdot 10^{-4}$
Physics	131652	475	10340	0,079	$1,6 \cdot 10^{-4}$

Table 1. Bundle statistics

3.2 Structural analysis

Figure 1 accounts for normalised occurrences of structurally classified bundles and their relative frequencies.

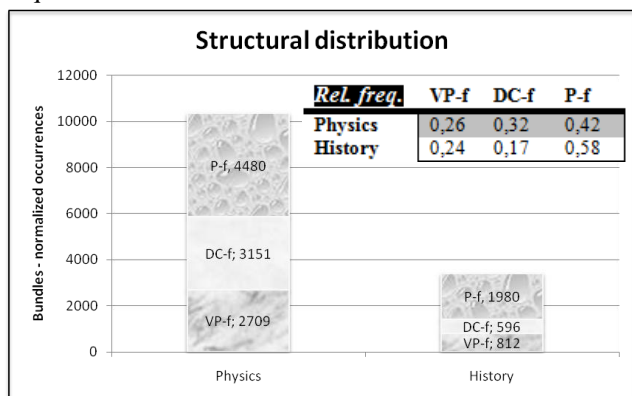


Figure 1. Structural distribution

The most common bundles from each category and subject are listed below. The numbers are normalised occurrences.

3.2.1 Samples from history

NP/PP-fragments

166 den industriella revolutionen (the industrial revolution)
 134 en del av (a part of)
 130 i början av (in the beginning of)
 114 i slutet av (in the end of)
 63 på så sätt (in that way)

VP-fragments

83 men det var (but it was)
 79 ta reda på (find out)
 51 det var inte (it was not)
 31 men det fanns (but there was)
 31 ta sig fram (take one's way)

Dependent clause fragments

59 för att få (to get)
 35 hade rätt att (was allowed to)
 35 ledde till att (resulted in)
 31 det gällde att

(it was a matter of)
 23 av dem som (of the ones that)

3.2.2 Samples from physics

NP/PP-fragments

284 med hjälp av (by means of)
 123 på grund av (because of)
 113 en del av (some of)
 111 på så sätt (in that way)
 82 på samma sätt (in the same way)

VP-fragments

84 det vill säga (which means)
 72 vad är det (what is it)
 68 ta reda på (find out)
 54 hur stor är (how big is)
 52 består av en (consists of a/one)

Dependent clause fragments

91 det beror på att (it is because of)
 86 det betyder att (it means that)
 77 vad som händer (what happens)
 72 så att den (in order for it to)
 66 lika stor som (as big as)

3.3 Functional analysis

The functional distribution of the 30 most prevalent bundles of the subjects respectively are presented in figure 2 below. The distribution is almost identical across the subjects, with about 70% being used for referential purposes and the remaining ones for discourse organization.

In Biber et al. (2004), the textbooks contained mostly referential expressions and subsequently discourse organizers and stance expressions to a lower extent. In the present case though, no stance expressions were discovered.

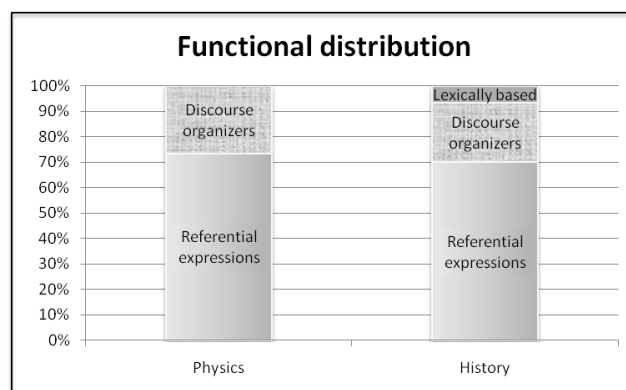


Figure 2. Functional distribution

Below are some sentences to exemplify the usage.

3.3.1 Samples from physics

Referential expressions

- På samma sätt kan det bildas gnistor när du kammar dig. (In the same way sparks can occur while combing one's hair.)

Discourse organizers

- Tänk dig ett järnfilspån som är en tiondels millimeter i diameter... (Imagine an iron filing that is a tenth of a millimeter in diameter)

3.3.2 Samples from history

Discourse organizers

- För att få in pengarna lade britterna tull och skatt på vissa varor som skulle till Nordamerika. (To bring in the money the Brits added customs and taxes to certain goods destined to North America.)

Referential expressions

- En del av vägen hade Bartolomeo Diaz redan kartlagt. (Bartolomeo Diaz had already mapped out a portion of the road.)

3.4 Comparison between the subjects

Figure 3 is a visualisation of the structures that make up the functions.

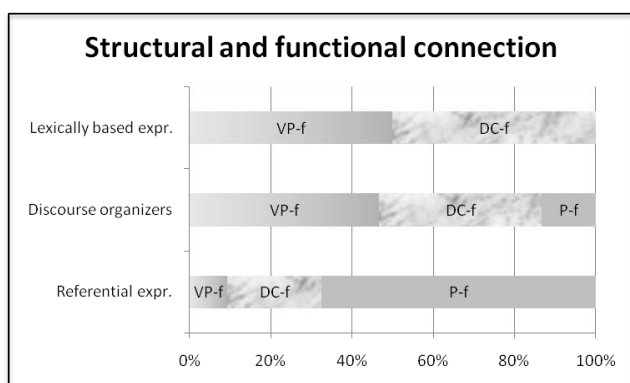


Figure 3. Structural and functional connection

The lexically based expressions and the discourse organizers are almost entirely built up by VP-f and DC-f. This is in fact the case in both subjects. In other words, the subjects do not differ in structural discourse management.

However, when looking at the referential expressions one finds a striking difference (see figure 4 below).

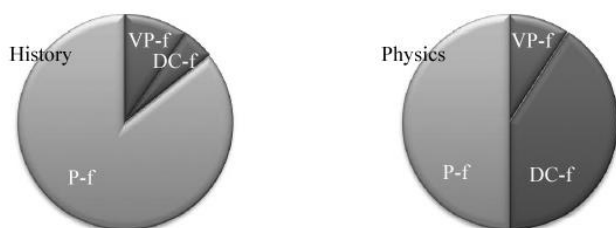


Figure 4. The structure of referential expressions

4. Discussion

From the results presented in table 1, it seems physics has a much higher bundle density, while at the same time a more varying language than history. A closer investigation however reveals that the bundle frequency highly differs among textbooks within the subjects. One of the physics books stands out by having very many bundles per word, while one of the history books stands out by having very few. Furthermore, it is important to bear in mind that the cut-off frequency applied in the identification process, if normalised with respect to the number of words, becomes

much higher in history than in physics⁴. All together, this study is not large enough to make any fixed conclusions about the possible differences in bundle density or variation between the two subjects.

The big tendencies seen figure 1 seem to be an equal use of VP-f:s, 25%, across subjects. Both history and physics have mostly P-f:s, with history being most extreme with almost 60% of all the bundles being P-f:s. DC-f:s are somewhat rare in history, yet more common in physics.

Looking at the bundles presented in 3.2.1 and 3.2.2, it is obvious that the P-f:s are by far most apt to being repeated throughout the texts. The five most common ones are also easily recognised as genuine lexical bundles. They can directly be understood as a whole, both conceptually and functionally. This strongly suggests that they are stored in one piece in our mind. Most of them contain both NP- and PP-fragments, i.e. nouns and prepositions in a fixed pattern. The prepositions in the bundles are not exchangeable to semantically close ones and the nouns have to be inflected as they appear. Consider, for example, *i början av* (in the beginning of), which cannot be switched for **på (on) början av*, **i början från (from)* or *#i starten (the start) av*. These all express the same lexical meaning, but would clearly qualify as unidiomatic language by native speakers of Swedish.

When it comes to VP-f:s and DC-f:s one cannot claim the bundles being genuine to the same extent. Some of them undoubtedly are, namely *ta reda på* (find out), *ta sig fram* (take one's way), *hade rätt att* (was allowed to) and *det vill säga* (which means). They are not as fixed as the P-f:s though, since the finite verb can be inflected; *tog reda på* (found out) and *det ville säga* (which meant) would also be genuine Swedish bundles.

The remaining bundles among VP-f:s and DC-f:s are a mixture of sequences containing collocates, e.g. **för att få** (to get) and **så att den** (in order for it to), phrasal verbs (**ledde till att** (resulted in) and **består av en** (consists of a/one) or fixed constructions, e.g. **lika stor som** (as big as) as well as sequences of mere frequent words without compound meanings, e.g. *men det var* (but it was), *av dem som* (of the ones that) and *hur stor är* (how big is).

As for figure 2 and the function of the most frequent bundles, in physics one can divine the subject's need for explaining cause-and-effect connections, e.g. *på grund av* (because of), *på så sätt* (in that way), *så att den* (in order for it to) and *det beror på att* (it is because of) and concepts, e.g. *det betyder att* (it means that) and *det vill säga* (which means). Moreover, bundles like *vad som händer* (what happens) and *ta reda på* (find out) indicate an exploring and analysing discipline. One can also predict the necessity to classify and relate terms and notions through *en del av* (some of), *lika stor som* (as big as) and *består av en* (consists of a/one).

⁴ A normalisation gives 30 and fully 50 times/million words for physics and history respectively.

Texts dealing with history also call for the ability to elucidate ideas, which shows through *på så sätt* (in that way), *ta reda på* (find out), *ledde till att* (resulted in) and *för att få* (to get). Other more subject-specific recurring sequences point to reasoning, e.g. *men det var* (but it was) and *det var inte* (it was not), ordering events chronologically, e.g. *i början av* (in the beginning of) and *i slutet av* (in the end of) or defining entities, e.g. *en del av* (some of) and *av dem som* (of the ones that).

Figure 4 shows that references in history by means of P-f:s are very much predominant, i.e. references to nominal, prepositional and comparative phrases. References to clauses, by DC-f:s, are rarely used in historic texts. Conversely, in physics quite a number of references are made through DC-f:s. Physical texts refer to and across clauses almost to the same extent as to phrases.

5. Conclusion

The most common bundles in physics have the function of classifying and relating concepts and entities, while history bundles help to discuss events and arrange them in time.

When looking at the relative frequencies of structural distribution across subjects, VP-f:s are equally spread, but history has more P-f:s and less DC-f:s than physics. The difference belongs to the referential function; history refers to phrases, i.e. discusses within the clauses, while physics also tends to make references across clauses. These findings are in accordance with the theoretical claims about the language made by Halliday and Martin (1993) of scientific and historical texts respectively.

The most surprising result was the small amount of bundles common to both subjects. Albeit every register previously has been reported to possess its own characteristic set of bundles, the textbook register was assumed to be more homogenous. In fact, textbooks from different disciplines cannot be generalized with respect to lexical bundles. Instead, more specialised investigations, both across subjects and form levels, are needed to better understand the distribution of lexical bundles across the genre. Extensive usability studies are also called for.

There are still many methodological questions in the field that need to be answered. What cut-off frequency works the best and is this consistent for identifying bundles of different lengths? Should the structural classification be context-independent? If so, it could easily be automated when working on a POS-tagged corpus.

The principal question for the near future to answer is definitely how to improve the identification process in order to find genuine bundles, i.e. the word boundaries between lexical bundles and surrounding arbitrary words.

Automated identification of discontinuous word combinations becomes the next concern.

References

Atkins, Sue B. T. and Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press: USA.

Attia, M., Toral, A., Tounsi, L., Pecina, P. and van Genabith, J. (2010). Automatic Extraction of Arabic Multiword Expressions. In: *Proceedings of the*

Multiword expressions: From Theory to Applications (MWE 2010), pp.19-27.

Biber, D. and Barbieri, F. (2007). Lexical bundles in university spoken and written registers. In: *English for Specific Purposes*, 26, pp.263-286.

Biber, D. and Conrad, S. (1999). Lexical bundles in conversation and academic prose. In: Hasselgard, H. and Oksefjell, S. (Eds.), *Out of corpora: Studies in honor of Stig Johansson* (pp.181-189). Amsterdam: Rodopi.

Biber, D., Conrad, S. and Cortes, V. (2004). *If you look at...: Lexical Bundles in University Teaching and Textbooks*. In: *Applied Linguistics* 25/3, pp.371-405. Oxford University Press.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. In: *English for Specific Purposes* 23, pp.397-423.

Cummins, J. (1980). The entry and fallacy in bilingual education. In: *NABE Journal*, IV(3). Norwood, NJ: Ablex.

Gardner, D. (2007). Validating the Construct of *Word* in Applied Corpus-based Vocabulary Research: A Critical Survey. In: *Applied Linguistics* 28/2, pp.241-265. Oxford University Press.

Halliday, M.A.K. and Martin, J. (Eds.). (1993). *Writing science: Literacy and discursive power*. Pittsburgh, PA: University of Pittsburgh Press.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. In: *English for Specific Purposes*, 27, pp.4-21.

Lindberg, I. and Johansson Kokkinakis, S. (2007). *OrdiL – en korpusbaserad kartläggning av ordförrådet i läromedel för grundskolans senare år*. ROSA-rapport 8. Gothenburg: Swedish institute.

Nekrasova, T. M. (2009). English L1 and L2 Speakers' Knowledge of Lexical Bundles. In: *Language Learning* 59:3, pp.647-486. University of Michigan.

Macken-Horarik, M. (1996). Literacy and learning across the curriculum: towards a model of register for secondary school teachers. In R. Hasan & G. Williams (Eds.), *Literacy in Society*, pp. 232-279. London: Longman.

McEnery, T., Xiao, R., Tono, Y. (2006). *Corpus-based language Studies: An advanced resource book*. London: Routledge.

Schleppegrell, M. J. (2004). *The language of schooling: a functional linguistics perspective*. Mahwah, NJ: Lawrence Erlbaum.

Simpson-Vlach, R. and Ellis, N.C. (2010). An Academic Formulas List: New methods in Phraseology Research. In: *Applied Linguistics* 31/4, pp.487-512. Oxford University Press.

Strunkyte, G. och Jurkunaite, E. (2008). *Written Academic Discourse: Lexical Bundles in Humanities and Natural Sciences*. Vilnius: Vilnius University.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Zhang et. al. (2006). Automated Multiword Expression Prediction for Grammar Engineering. In: *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pp.36-44. Sydney: Association for Computational Linguistics.