Journal of Social Sciences

Original Research Paper

Sequential Patterns: A new Corpus-Based Method to Inform the Teaching of Language for Specific Purposes

Dominique Legallois and Anne Prunet

Crisco, University of Caen, Caen, France

Article history
Received: 07-05-2015
Revised: 19-05-2015
Accepted: 19-05-2015

Corresponding Author:
Dominique Legallois
Crisco, University of Caen, Caen, France
Email: dominique.legallois@unicaen.fr

Abstract: In this study, we present an original study that aims to show how pedagogically important lexico-grammatical patterns, which are typical of certain genres, can be identified and taught at every educational level, in particular for learners of French for Specific Purposes. These patterns, called Sequential Patterns, constitute a more powerful paradigm than lexical bundles or "P-frames" because they combine different levels of abstraction (word forms, lemmas, POS tags). As they are typical of a textual genre, some of these items reveal the abstract phraseological dimension of texts.

Keywords: Sequential Patterns, Teaching of Language for Specific Purposes, Lexical Bundles, P-frames, Phraseology

Introduction

The goal of this paper is to present a new approach to language description based on corpus investigation techniques. This approach consists in identifying the lexico-grammatical units that are characteristic of genres. These units are called "sequential patterns". They combine different levels of abstraction (word forms, lemmas, POS tags), which provide more or less generic patterns. For instance: *To the N; he V the N of his ADJ N; the N V a/an ADJ N for the N; the N is that*, etc. Such patterns, as will be argued in this study, represent a link between lexis, grammar, usage and texts. Some of them should be regarded as phraseological or formulaic units, despite their abstract nature. For example, the French sequential pattern *le N comme N de le N de le N* (the N as N of the N of the N) is typical of the genre "philosophical essays":

1 la question de l'être, comme question de la possibilité du concept (Derrida)
Lit ("lit." means that we provide literal translations of French examples). The question of being, as the question of the possibility of the concept

2 la révolution comme condition de la réalisation de la jouissance (Onfray)
lit. the revolution as the condition for the realization of the enjoyment

This abstract pattern is under-used in other genres such as political debate or oral conversation (The term genre is used here to refer to a recurrent social practice, characterized by a set of conventional and organized constraints on production and interpretation). It invariably expresses the characterization of a (philosophical) concept (*question of being, revolution*). In this way, a relatively fixed form, although generic in nature, carries out a constant function: It is a reproducible syntagmatic unit, which has a relative syntactic and semantic stability. It constitutes a way of speaking, a manner in which words or phrases are used in a particular genre. For these reasons, we consider this pattern as a phraseological unit. Phraseology is thus defined in this study as "the preferred way of saying things in a particular discourse" (Gledhill, 2000).

We hypothesize that identifying the significant sequential patterns in a given genre is useful and important for language teaching because it can facilitate both text recognition and production. Since the process of learning a language must be seen as the process of acquiring the relevant patterns which codify the conventions of use of language in context, we assume that sequential patterns are part of these relevant units and as such, must be used as the basis for materials design and curriculum development. Teachers should draw attention to Sequential Patterns in class and use them as the basis for explicit instruction.

Science Publications

In the framework of this new approach to language description, this paper aims to contribute to our understanding of the efficacy of genre-based teaching for building learners' genre knowledge and improving their L2 writing and reading abilities. The focus of this paper will be on French for Specific Purposes (FSP) and the teaching of written scientific texts.

The article begins with a review of previous inductive corpus research on lexico-grammatical patterns: Lexical bundles and P-frames. Their importance for language teaching is highlighted as well as their limits. In the following section, we describe the method used for the extraction of Sequential Patterns and present an empirical investigation of Sequential Patterns that are typical of scientific discourse. The corpus investigated (TAL corpus) consists of Natural Language Processing papers (The TALN proceedings corpus is a subset of the scientific articles presented at the Traitement Automatique des Langues Naturelles (TALN) and Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL) conferences from 2007 to 2013. This corpus consists of 586 articles totalling about 2 million words. It is available at http://redac.univ-tlse2.fr/corpus/taln_en.html. For the present study, we used part (986 049 words) of the whole corpus): We identify Sequential Patterns of unusual frequency in this genre in comparison with a suitable reference corpus (political discourse and critical reviews). This method highlights the typical structures of the scientific genre, but also the constructions that can cause problems from a teaching perspective.

The aim of the third section is twofold: (1) To demonstrate how the phraseological dimension of a text genre can be uncovered; (2) to present the key steps involved in creating inventories of Sequential Patterns, according to their location in texts.

The approach advocated here is consistent with both Construction Grammar (Goldberg, 1995) and Pattern Grammar theories (Hunston and Francis, 2000); it shares with them the view that phrasal constructions which consist of grammatical, lexical, semantic and pragmatic relations are the primary units of meaning. There is no clear-cut division between regular and phraseological expressions. As a result, these expressions can be simply taught and learned as whole units.

Previous Research on Lexico-Grammatical Patterns in Language Teaching

Corpus Linguistics

Over the last twenty years, corpus linguistics has become one of the fundamental methods for analyzing languages. It is used in a wide variety of applied settings: Genre analysis, lexicography, discourse analysis, stylistics, grammar, description of specialized texts, language teaching, translation, phraseology-to mention just a few. It is characterized by four main features (Biber *et al*., 1998): (1) Corpus linguistics is, by definition, empirical: Language patterns are observed in real (spoken or written) texts; (2) these texts are representative samples of language and are stored as electronic corpora; (3) the analysis relies on computer software or programs to identify and count linguistic patterns; (4) the analysis depends on both quantitative and qualitative techniques.

With the development of corpus linguistics, electronic corpora and software tools, there has been increasing interest in the ways in which phraseology or formulaic language (Wray, 2002) -the tendency for words to appear in some environments more than others-can be identified in large text corpora. While this investigation can be conducted with a simple analysis of concordance lines, the limitation of this approach is that it is not inductive: The linguist must make a pre-selected choice of a particular word. Fortunately, several different approaches have been developed to uncover new linguistic constructs through inductive corpus analysis. In this section we briefly present two methods: Lexical bundles (or n-grams) and P-frames (or collocational frameworks). Both these methods are inductive; they do not require human intervention, such as the selection of patterns, or searches for specific units. They are said to be “corpus-driven” since the computer operates on raw texts, without any extra information.

Lexical Bundles

Lexical bundles (or n-grams) are frequently recurring sequences of words in texts. They are strings of contiguous words that constitute a phrase or a pattern of use. They are automatically extracted from corpora. These units give insight into important aspects of the phraseology used by writers in specific genres. For example, the following recurring sequences of three or four words are frequently used in English academic writing (Biber, 2009): *The fact that, at the same, seems to be, the same time, as a result of, on the other hand, a member of the, it is important to, I'd like to*, etc.

The Table 1 shows a selection of the top thirty 4-grams of the TALN Corpus.

The taxonomy of lexical bundles proposed by Hyland (2008) can be applied here to categorize these units. This taxonomy is based on Halliday’s metafunctions: Ideational, textual and interpersonal (Halliday, 1994).

Research-oriented lexical bundles (ideational metafunction): The units are used to express the activities and experiments of the investigator:

- Procedure: *La prise en compte*, etc.
- Quantification: *Un certain nombre de*, etc.
- Topic: *Traitement automatique des langues, état de l'art*, etc.

Table 1. Lexical clusters in the TAL corpus

c'est à dire (that is to say)	un point de vue (a point of view)	il est possible de (it is possible to)
dans le cadre de (in the framework of)	c'est le cas (it is the case)	d'un ensemble de (of a set of)
dans cet article nous (in this article we)	la prise en compte (the consideration of)	état de l'art (state of the art)
à l'aide de (with the help of)	et d' autre part (and, on the other hand)	au niveau de la (at the level of the)
dans un premier temps (in a first time)	il n' y a (there is no)	comme nous l'avons (as we have)
à partir d'un (from a)	un certain nombre de (a number of)	à partir de corpus (from a corpus)
il s' agit de (it is a question of)	traitement automatique des langues (Natural Language Processing)	il n'est pas (it is not)
du point de vue (from the point of view)	d'un point de (from a point of)	la recherche d'information (information retrieval)
en ce qui concerne (in respect of)	à partir d'une (from a)	dans la mesure où (insofar as)
dans le cas de (in the case of)	sous la forme de (in the form of)	sur la base de (of the basis of)

Text-oriented lexical bundles (textual metafunction): The units are used to structure the text:

- Structuring signals: *Dans cet article nous, dans un premier temps, et d'autre part*, etc.
- Framing signals: *Dans le cas de, dans le cadre de, d'un point de vue, du point de vue*, etc.
- Transition signals: *En ce qui concerne*, etc.

Participant-oriented lexical bundles (interpersonal metafunction): The units focus on the writer's and / or the reader's attitude or evaluation:

- Writer's attitude: *Nous avons montré que*, etc.
- Evaluations: *Il est possible*, etc.

One approach to using lexical bundle analysis for teaching purposes is to compare and contrast the lexical bundles produced by students in their writings with the lexical bundles produced in professional prose. Scott and Tribble (2006) carried out a contrastive analysis of this kind. Their study showed how the performances of advanced apprentice undergraduate writers in literary studies (Polish MA English Philology students) contrast with those of authors of published research articles in the same field. For example, Scott and Tribble noted the low use of the "anticipatory it" four-word lexical bundle in the students' texts (*it is hard to, it is possible to, it is true that*, etc.). They attributed this low use to the fact that there is less evaluation in the students' dissertations (or that evaluation is expressed with other forms).

Another study on lexical bundles in academic writing, by Chen and Baker (2010), compared the frequently-used word combinations in a corpus of published academic texts and two corpora of student academic writing (one L1, the other L2). They concluded that both in L1 and L2 corpora, formulaic expressions increase with writing proficiency-in other words, the number of lexical bundles increases with advancing writing proficiency, which is the case both for the range of lexical bundles used (types) and the overall occurrence of lexical bundles (tokens). The study demonstrates that prefabricated lexical patterns are not strategically used by learners, at a basic level, in order to facilitate writing, but are the result of an increasing mastery of textual norms.

Such investigations have some pedagogical implications. They argue in favor of a better integration of lexical bundles in teaching materials and call for an increased pedagogical focus on these phraseological units.

P-Frames

P-frames (Fletcher, 2003) or collocational frameworks are sequences made up of high-frequency function words as fixed elements co-occurring with variable internal lexical slots, e.g. _be + * + to; a/n + * + of + the; many + * + of; too + * + to_ (the symbol * is used to represent a gap of one word). For example, the collocational frameworks _too + * + to_ co-occurs most commonly with the fillers *young (too young to), little, stupid*, etc.; _a + * +of_ his frequent realizations such as *a lot of, a number of, a couple of, a series of, a variety of*, etc. Some units are more "lexical", since they are determined more directly by lexical words. For example (Stubbs, 2007): _Plays a * part in_ (_* = large, significant, big, major, vital, essential, key, central, full, great, prominent_).

P-frames are more flexible and schematic than lexical bundles, since they provide systematic groupings of lexical bundles, which vary in only one position.

These discontinuous patterns were first described by Renouf and Sinclair (1991). Their approach was corpus-based, but not inductive since they pre-selected a small number of collocational frameworks that they had noticed and then studied them in a corpus to identify the lexical words that filled the variable slots. Recently, automatic and unsupervised extraction of P-frames from large corpora has been made possible for non-computer specialists with software such as kfNgram (http://www.kwicfinder.com/kfNgram/kfNgramHelp.html; note also that the PIE (Phrase In English) database provides a powerful interface for studying N-grams and P-Frames (http://phrasesinenglish.org/)). The Table 2 gives a selection of the first thirty 4-frames of the TALN corpus:

Table 2. First thirty 4-frames of the TALN corpus

de la * de (of the * of)	sur la * de (on the * of)	des * et des (some * and some)
la * de la (the * of the)	à la * des (to the * of the)	de la * du (of the * of the)
dans le * de (in the * of)	de la * et (of the * and)	dans * cadre de (in * framework of)
le * de la (the * of the)	à * de la (at * of the)	dans cet * nous (in this * we)
de * et de (of * and of)	il est * de (it is * of)	dans * article nous (in * article we)
de la * des (of the * of the)	dans un * temps (in a * time)	dans les * de (in the * of)
à la * de (to the * of)	dans la * de (in the * of)	pour la * de (for the * of)
de * de la (of * of the)	au * de la (at the * of the)	des * de la (some * of the)
les * de la (the * of the)	sur les * de (on the * of)	de * nous avons (of * we have)
les * et les (the * and the)	sur le * de (on the * of)	à partir * corpus (from * corpus)

From this table, it can be seen that collocational frameworks are extremely abstract units. There is no categorization or specification of the slot "*", which makes their application to pedagogical practice highly problematic: It is obvious that they are too complex and schematic to be taught in the classroom.

In addition, some P-frames can be said to be homogeneous: The slot is filled with items belonging to the same lexical domain (or having the same function). For example:

> *il est * de* (it is * of)
> In which the asterisk represents an evaluative adjective: Intéressant, aisé (easy), *notable, possible, important, difficile...*

In this case it can be said that words which regularly occur with similar patterns tend to share aspects of meaning. One could therefore consider *il est * de* a phraseological P-frame. However, many P-frames are more volatile and heterogeneous. For example:

> *dans le * de* (in the * of)
> in which the slot can be filled with very different words:
> *dans le domaine de + N* (in the domain of) – the complement is a noun;
> *dans le but de + INF* (in the aim of) – the complement is an infinitive.

The pedagogical use of this P-frame is therefore very complicated and its relevance is questionable.

To our knowledge, the P-frames methodology has never been applied in the field of language teaching, except for one study by Römer (2009), who compared the P-frames from three corpora (She also analysed keywords and lexical bundles):

- A collection of 45 essays written in English by upper-level university students at the universities of Cologne and Hanover, in English linguistics and English Literary Studies
- A subset of 91 English (language and literature) and Linguistics papers from the Michigan Corpus of Upper-level Student Papers (by native students)
- 30 published research articles from the field of Linguistics in the Hyland Corpus, written by native experts

The comparison revealed that:

- There was substantial overlap of 3 and 4 P-frames between non-native and native students
- Non-native and native students differ in similar ways from expert academic writers

The conclusion of this study leads to an important question:

> "The information that we obtained seems to indicate that, when we deal with advanced-level academic writing, we actually move beyond the native/non-native distinction and that, in this context, experience or expertise is a more important aspect to consider than nativeness" (Römer, 2009)

Such a counterintuitive conclusion has significant implications for the very nature of the pedagogical practice and its specific characteristics.

Discussion

Lexical bundles and P-frames are useful methods for analyzing languages and texts. In addition, corpus analysis has proved to be relevant and efficient for the study of phraseology and for pedagogical material design (although adjustments are still needed for P-frames). However, the question of the granularity of the linguistic forms inevitably arises: On the one hand, lexical bundles are too specific. They are unmodified repetitions of pieces of texts. On the other hand, P-frames are too schematic. They are complex amalgamations of many different forms.

It is clear that teaching materials need to provide both conventionalized and fixed expressions and semi-specific sequences that can be realized in different ways. These semi-specific (or semi schematic) sequences must be of a sufficient granularity to illustrate the core pattern of the unit and the typical lexico-grammatical realizations of this unit. We hypothesize that Sequential Patterns may have the potential to address this need.

Sequential Patterns

Procedure

To a certain extent, it can be said that the Sequential Patterns method is a kind of mix between lexical bundles and P-frames since Sequential Patterns are patterns that combine different levels of abstraction (word forms, lemmas, POS tags). Therefore, Sequential Patterns have a multidimensional nature (Quiniou *et al*., 2013). To illustrate the method of extraction with a concrete example, let us take this sentence from the corpus:

3 Il sera particulièrement intéressant de comparer le pouvoir discriminant de dépendances syntaxiques comparables en fonction de la paire de langues considérée lit. It will be particularly interesting to compare the discriminating power of comparable syntactic dependencies in function of the pair of languages considered

Stage 1

The sentence is annotated in a hybrid manner:

- Invariable words (pronouns, determiners, adverbs, conjunctions, prepositions) are not pos-tagged. In example 3, this applies to the pronoun *il* (it), the preposition *de* (of), the prepositional locution *en fonction de* (as a function of), the determiner *le* (the)
- Very frequent lexemes are lemmatised (the verb *être*-to be)
- Others lexemes are pos-tagged (The Cordial tagger was developed by Synapse Développement (www.synapse-fr.com)) (ADV = adverb; ADJ = adjective; N = noun; PRES = present participle, PASS = past participle, NP = proper noun, INF = infinitive, etc.)

The annotation of the sentence 3 is:

- il être ADV ADJ de INF le N PRES de N ADJ ADJ en fonction de le N de N ADJ

The whole corpus was annotated in this way.

Stage 2

All the n-grams (from 4-grams to 8-grams) are automatically extracted from the corpus (for example *il être ADV ADJ de INF*; *le NC PRES de NC ADJ*; *en fonction de le NC de NC ADJ*). The result is a list such as:

51	un N de le N
48	le N ADJ le N
47	N ADJ le N de
45	de le N ADJ de
44	dans le N de le N
42	N le N de N
41	N ADJ à le N
40	N PASS par le N
40	ADJ de le N ADJ
…	…………………..

Stage 3

Chunks of texts matching the n-grams are identified by a "return to the text". Note that not all the Sequential Patterns are interpretable. For example, *N le N de N* is not a coherent pattern because it does not correspond to an interpretable phrase (e.g., terme l'algorithme de recherche-lit. term the algorithm of research). No coherent patterns are removed.

Below, we give two examples of Sequential Patterns in the TAL corpus:

• **chaque N être PASS comme ADJ** (lit. each N to be past-participle as ADJ)

4 chaque partie est vue comme close
Lit. each part is seen as closed

5 chaque candidature est identifiée comme pertinente
Lit. Each candidacy is identified as relevant

6 chaque terme est considéré comme syntaxique
Lit. Each term is considered as syntactic

• **nous avoir mettre en N un N** (lit. we have put in N DET N)

7 nous avons mis en évidence un algorithme de recherche
Lit. we have highlighted a search algorithm

8 nous avons mis en place deux types de contrainte
lit. We have implemented two types of constraint

9 nous avons mis en œuvre des règles de résolution
lit. We have implemented resolution rules

As the examples above show, Sequential Patterns constitute a more powerful paradigm than lexical bundles, since lexical bundles can be seen as specific instances of Sequential Patterns. Sequential Patterns can be compared with the Patterns investigated by Pattern Grammar theory (Hunston and Francis, 2000); like Patterns (for example, V N by N), Sequential Patterns contain a mixture of lexical items (mostly prepositions, conjunctions, frequent verbs, etc.) and abstract elements (such as nouns, adjectives, verbs, etc.). The main difference is that the former are automatically generated by computers.

Specific Sequential Patterns

Specific (or 'keyword') Sequential Patterns can be calculated. Specific Sequential Patterns are the Sequential Patterns that occur significantly more often in the corpus than one would predict. In another words, a Specific Pattern is a pattern that is particularly frequent (a 'positive' keyword) or infrequent (a 'negative' keyword) in a corpus in comparison to its frequency in a reference corpus. By comparing the frequency of a sequential pattern in the TAL Corpus to the frequency of the same pattern in the Reference Corpus, it is possible to ascertain whether the pattern appears more frequently in the TAL Corpus than predicted, through the underlying statistical analysis based on the exact probabilities of the hypergeometric distribution (Lafon, 1984). The Reference Corpus includes political debates in the French National Assembly (999 172 words) and book, film and music reviews ('evaluative texts': 793 055 words) (From http://www.avoir-alire.com/). The results were calculated by the statistical software R (http://www.r-project.org/).

Three of the TAL Corpus Specific Sequential Patterns are given below:

- **le N de ce N être de INF**

10 L'objectif de cet article est de présenter...
lit. The goal of this paper is to present

11 Le but de cette expérience est de montrer...
lit. The aim of this experiment is to show

12 Le propos de cet article est de présenter...
lit. The point of this paper is to present

- **ce N V de INF**

13 Cet article propose de combiner les résultats de plusieurs moteurs de traduction automatique issus de systèmes de natures différentes
lit. This article proposes to combine the results of several machine translation engines from systems of different natures

14 Cette heuristique a pour but de déterminer l'hyperonyme du sujet de l'article
lit. This heuristic aims to determine the superordinate theme of the paper

15 Cette typologie permet de proposer une première interprétation
lit. this typology allows to propose a first interpretation

These two constructions have an obvious role in the structuring of texts. They are text-oriented.

- **nous avoir PASS de INF**

16 Nous avons décidé de structurer les objets qui composent notre représentation dans une ontologie Lit. We have decided to organize the objects that make up our representation in an ontology

17 Nous avons choisi de représenter le réseau avec des déclarations RDF lit. We have chosen to represent the network with RDF statements.

18 Nous avons proposé d'utiliser des techniques d'EI
Lit. We have proposed to use the techniques of EI

This Specific Sequential Pattern conveys the writer's attitudes. It is participant-oriented.

Implications for Language Teaching

The implications of some of these Sequential Patterns for teaching French for specific purposes are briefly presented in this section. We focus here on two families of Sequential Patterns which present considerable difficulty for learners.

First Example: Complex Noun Phrases

Genres relate to social-cultural practices; they are determined by communicative settings and reflect different cognitive representations. They develop specific linguistic forms. These forms can be considered as phraseological because they reflect the manner in which words or phrases are used in a particular genre.

For example, compared to the debates in the French National Assembly and to evaluative texts, scientific papers greatly favor heavy nominal groups and nominalizations, in which the head noun is typically accompanied by pre-modifiers such as adjectives or nouns and/or by post-modifiers such as prepositional phrases.

- **le N de N de N ADJ**

19 l'objectif final est de disposer d'un analyseur qui facilite l'annotation de corpus de définitions terminographiques
lit. The ultimate goal is to develop an analyzer that facilitates the annotation of corpora of terminographical definitions

20 Nous n'avons pas encore évalué les méthodologies d'extraction de relations syntagmatiques
lit. We have not yet assessed the methodologies for extraction of syntagmatic relations

- **un N de le N de un N de N**

21 Dans cet article nous présentons une évaluation et une analyse des résultats d'une méthode de réordonnancement de réponses pour un système de questions-réponses
lit. in this study we present an evaluation and an analysis of the results of a method of reordering responses for a system of questions-answers

While mastering noun phrase complexity and use can be analyzed as clear evidence of satisfactory proficiency, FSP textbooks do not exploit this complexity in any systematic or meaningful way, although it is clear that FSP students often have problems understanding or producing these structures. Several exercises or activities can be proposed, such as:

- Distinguishing the relevant head noun from pre-modifiers and post-modifiers
- Producing nominalizations from verbal phrases
- Building lengthy sentences by increasing the complexity of noun phrases

Second example: the "se V" construction

A transitive verb is called reflexive when its action returns upon the actor, in other words, when the subject and object are identical:

22 je m'approche de la fenêtre
lit. I move myself closer to the window. I move closer to the window

However, reflexive verbs can take patient subjects in a "se V" construction. A "se V" construction is a grammatical structure in which an a priori transitive verb is linked to the "reflexive" pronoun *se* (itself). The pronoun carries out the agentivity of the verb and the subject is the patient of the process. The result is that, contrary to the example *je m'approche de la fenêtre*, the sentence has a passive meaning. For example:

23 Ce problème s'explique facilement
lit. This problem explains itself easily. This problem is easy to explain

This construction occurs very frequently in scientific writing. From a grammatical point of view, the construction competes with the active voice + an impersonal pronoun (*on explique facilement ce problème*/one may easily explain this problem) or with a typical passive form (*ce problème est expliqué facilement*/this problem is easily explained). From a discursive perspective, however, the "se V" construction is the norm, whereas the active voice with the impersonal pronoun *on* and the passive voice are infrequent.

The specificity of the construction should not be taught exclusively as a purely grammatical device, but also as a discursive form with its semantic functions. As such, Sequential Patterns are helpful. They provide us with all the configurations in which the "se V" construction plays a role:

- **N se V par le N**

Is a Sequential Pattern used:

- To express a result

Depending on the verb, A is the result of B or B is the result of A:

24 La génération complète d'un document se fait par la production successive de tous les groupes lit. The complete generation of a document is made (se fait = makes itself) by the successive production of all the groups

Interpretation "A is the Result of B": The complete generation of a document is the result of the successive production of all the groups:

25 L'analyse qualitative se traduit par la mise en place d'un schéma d'annotation
lit. the qualitative analysis is reflected (*se traduit* = translates itself) in the establishment of an annotation scheme

Interpretation "B is the Result of A": The establishment of an annotation scheme is the result of the qualitative analysis:

- To express an explanation

26 une grande part de la perte de performance s'explique par le bruit contenu dans le corpus
lit. much of the performance loss is explained (s'explique = explains itself) by the noise contained in the corpus.

- to characterize the subject

27 Les résultats […] se distinguent par le fait d'utiliser une combinaison de multiples espaces de forte dimensionnalité
Lit. The results [...] are distinguished by the fact of using a combination of multiple high dimensional spaces

28 Ce qu'on appelle état se caractérise par l'absence de changement
lit. What we call state is characterized by the absence of change

The "se V" construction is also used in other Sequential Patterns:

• se V sur le NC de NC

This Pattern expresses the basis of an analysis:

29 l'analyse morphologique s'appuie sur les règles de formation des mots
lit. The morphological analysis is based on the word formation rules
30 Notre première approche se fonde sur l'écriture de règles
lit. Our first approach is based on the writing of rules
31 L'approche classique se base sur l'alignement de mots
lit. The classical approach is based on the alignment of words

• se V en ADJNUM (Numeral Adjective) N

This Sequential Pattern fulfills a "prospective" function:

> "Prospection occurs where the phrasing of a sentence leads the addressee to expect something specific in the next sentence" (Sinclair, 2004):

32 Notre chaîne de traitement se décompose en trois étapes
lit. Our processing chain consists of three steps.
33 La génération du groupe de propositions […] se déroule en deux phases
lit. The generation of the group of proposals [...] takes place in two phases.

These examples show that a grammatical phenomenon should be simultaneously taught with the discursive or textual functions that it concretely helps to express.

Distribution of Sequential Patterns and Lexical Bundles

Distribution of Sequential Patterns

In this section, we investigate the distribution of lexical bundles and Sequential Patterns in the scientific paper (Römer (2010)): In which parts of the text do these units preferentially appear? In the abstract, in the introduction, in the first part (first section), in the last part, or in the conclusion?

We analyzed thirty papers in the TAL Corpus. Applying the statistical method described above, we identified the units significantly attracted by the five parts. The results with frequencies below are presented in the different categories:

Abstract

- Lexical bundles: *Cet article présente* (18)*; dans cet article nous* (16)
- Sequential Patterns: *Dans ce N nous V un N* (15)*; V un N pour INF*(10)*; le N de N de N ADJ* (29)*; nous V en N le N* (11)*; un N ADJ de le N* (40)

Introduction

- Lexical bundles: *Cet article est* (25)*; dans le cadre de* (18)*; le domaine de* (16)*; dans cet article* (52)*; en termes de* (12)*; dans la section* (28)*; du point de vue* (20)*; dans cet article nous* (22)
- Sequential Patterns: *Dans le N NCMIN nous V* (43)*; le N en N de un N* (30); *N à N de N NP* (25)*; nous V le N de* (61)

1st part

- Lexical bundles: *A l'aide de* (22)*; comme par exemple* (21)*; dans la phrase* (15)*; de ce type* (17)*; le plus souvent* (13)*; l'analyse de* (10)*; nous présentons dans* (13)*; tout d'abord* (23)
- Sequential Patterns: *Le N entre N ADJ* (34)*; ne être pas ADJ de* (28)*; V à le N de INF*(18)

Last Part

- Lexical bundles: *Le nombre de* (65)*; les résultats de* (40)*; dans le texte* (25)*; par rapport à* (30)*; dans le tableau* (25)*; il s'agit de* (26)*; les résultats obtenus* (28)*; par rapport au* (27)
- Sequential Patterns: *Avec NP V N et NP V* (48)*; le N de chaque N* (37)*; il être ADJ de INF que* (39)*; et NP V N* (33)*; PASS à le N NCMIN* (53)*; N par N à le N* (66)*; le N parmi le N* (31)*; N avec le N NP (33); avoir être PASS à le N de* (36)*; que le N de N êtr*e (35)

Conclusion

- lexical bundles: *Dans cet article nous avons* (16)*; nous avons présenté* (18)
- Sequential Patterns: *N pass sur le N* (30)*; avoir PASS que le N* (37)*; ADJ à la N de* (48)*; le N PASS dans ce N* (40)*; N ADJ V un N* (48)*; N plus ADJ de N* (20)

We will discuss a few Sequential Patterns in relation to the sections of the paper in which they preferentially occur.

Section: Introduction

• nous V le NC de

34 Puis, nous exposons les conditions de nos expériences, en détaillant les particularités du corpus qui a servi à les mener
lit. then, we describe the conditions of our experiments, detailing the specificities of the corpus that was used to conduct them

35 Après avoir décrit notre système et le système de prédiction que nous avons utilisé, nous présenterons les résultats de deux expérimentations
lit. after describing our system and the prediction system that we used, we will present the results of two experiments

This Sequential Pattern illustrates the lexical (to describe/to present), temporal (present/future) variation of a discourse organizer. More precisely, the pattern is linked to Swales' move 3, "Occupying the niche" (that is to say, introducing the current research study in the context of previous research) and illustrates step 3, "Indicating the research paper structure" (Swales, 1990; 2004).

Section: First Part

- **ne être pas ADJ de INF**

36 notons qu'il n'est pas nécessaire d'encoder toutes les UT de manière identique
lit. note that it is not necessary to encode all the TU in the same manner

37 Néanmoins, la représentation que nous proposons n'est pas exempte de critiques
lit. nevertheless, the representation that we propose is not free from criticism

38 Dans cet article, il ne sera pas nécessaire de préciser si le texte est en format papier ou électronique
lit. in this article, it will not be necessary to specify whether the text is in paper or electronic format

This sequential Pattern is used to express a denial against the putative addressee, i.e., against beliefs which the writer assumes that at least some members of his or her audience will hold. Denials anticipate some possible misunderstanding or misconception on the addressee's part. It is only logical that such a Sequential Pattern is mainly found in the first part of a paper.

Section: Last part

- **le N entre N ADJ et N**

39 La différence entre contexte processif et contexte artefactuel est plus aisée à reconnaître que celle entre contexte dynamique et contexte statique.
lit. The difference between litigious context and artifactual context is easier to recognize than the difference between dynamic context and static context.

40 le parallélisme entre polysèmes logiques prototypiques et noms d'action ambigus n'est pas total
lit. The parallelism between logical prototypical polysemes and nouns of ambiguous action is not total

This Sequential Pattern contributes to the evaluation of the comparison of two objects. Its location at the end of the paper, in the last part where the author evaluates the findings, makes perfect sense.

Section: Conclusion

- **avoir PASS que le N**

41 Nous avons montré que la navigation dans un texte peut être modélisée à l'aide du langage Sextant
lit. We have shown that navigation in a text can be modeled using the Sextant language

42 Notre étude a établi que le meilleur de ces modèles est un classifieur SVM
lit. Our study has established that the best of these models is a SVM classifier

This Sequential Pattern exemplifies a formulation in which a special type of reporting verb is used (show, find, demonstrate, etc.) in the past tense. Propositions are construed by the authorial voice as correct, valid, undeniable or otherwise maximally warrantable. This Sequential Pattern is a key function in the rhetorical presentation of research since it contributes to signalling the main conclusions to be drawn from the study.

- **le N permettre de INF**

This Pattern with the verb *permettre* (to allow) encodes an "enablement" relation (Jordan, 1998), which is a special type of causal relation. The enablement relation increases the ability to perform an action. The Pattern expresses the work to be done in the near future. This future can be signaled in several ways: Either by the future simple tense:

43 Dans un second temps, l'extraction de relations entre entités nommées, [...], permettra de découvrir les groupes de gènes surexprimés
lit. in a second step, the extraction of relations between named entities, [...], will allow to discover the groups of overexpressed genes

Or by an infinitive expressing a perspective-for example, the verb *envisager* (to consider):

44 En conclusion, la méthodologie proposée permet d'envisager l'exploitation des marqueurs de thèmes pour les systèmes de segmentation automatique des textes
Lit. In conclusion, the proposed methodology allows to consider the exploitation of the markers of themes for automatic text segmentation systems.

All these Sequential Patterns are ready-made units appropriate for a particular situation. More exactly, they are primed for use in textual organization (Hoey, 2005). They have an association with one particular part of the text, such as Introduction, First Part, Conclusion, etc. The above analysis shows how a text-type specific inventory of abstract phraseological units together with their variation, functions and textual distribution can help FSP students to know which phrases to use and when and where to use them in a text.

Conclusion

The main purpose of this study has been to explore the extent to which phraseology contributes to academic writing by identifying not only lexical bundles and P-frames, but also Sequential Patterns. From a pedagogical perspective, our approach has several advantages:

- Patterns are automatically discovered rather than manually selected. The method gives new insights into typical lexico-grammatical structures of a textual genre that are not available through introspection or intuition
- Focusing just on lexical bundles would not give students access to the full range of formulaic units that are regularly used in academic writing. As they are more variable and schematic than lexical bundles, Specific Sequential Patterns help to determine the extent of the phraseological tendency of language
- Lists of Sequential Patterns can be made and used to help FSP course designers establish relevant teaching materials; they provide teachers and learners with conventional linguistic forms and phraseological units used for text structuring
- Since the Sequential Patterns are classified according to their functions in discourse, it is possible for users to access the lists based on what they wish to convey in the text they are composing
- Sequential Patterns are not distributed randomly across texts; they are connected to their structures and associated with particular textual positions

Teaching phraseology is a complex phenomenon that must be approached from several complementary perspectives for a full understanding. In this respect, we hope that our study may contribute to this goal.

Author's Contributions

All authors equally contributed in this work.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Biber, D., 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. Int. J. Corpus Linguist., 14: 275-311. DOI: 10.1075/ijcl.14.3.08bib

Biber, D., S. Conrad and R. Reppen, 1998. Corpus Linguistics: Investigating Language Structure and Use. 1st Edn., Cambridge University Press, Cambridge, ISBN-10: 0521496225, pp: 300.

Chen, Y.H. and P. Baker, 2010. Lexical bundles in L1 and L2 academic writing. Lang. Learn. Technol., 14: 30-49.

Fletcher, W., 2003. Exploring words and phrases from the British national corpus.

Hyland, K., 2008. As can be seen: Lexical bundles and disciplinary variation. English Specific Purposes, 27: 4-21. DOI: 10.1016/j.esp.2007.06.001

Gledhill, C., 2000. Collocations in Science Writing. 1st Edn., Gunter Narr, Tubingen, ISBN-10: 3823349457, pp: 432.

Goldberg, A., 1995. Constructions: A Construction Grammar Approach to Argument Structure. 1st Edn., University of Chicago Press, Chicago, ISBN-10: 0226300854, pp: 260.

Halliday, M., 1994. An Introduction to Functional Grammar. 2nd Edn., Edward Arnold, London, ISBN-10: 0340574917, pp: 419.

Hoey, M., 2005. Lexical Priming: A New Theory of Words and Language. Routledge, London, ISBN-10: 0415328632, pp: 196.

Hunston, S. and G. Francis, 2000. Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English. 1st Edn., John Benjamins, Amsterdam, ISBN-10: 9027222738, pp: 275.

Jordan, M., 1998. Enabler-enablement relations in English. Forum of the Linguistic Association of Canada and the United States (LACUS), 24, Toronto, Canada.

Lafon, P., 1984. Dépouillements et Statistiques en Lexicométrie. 1st Edn., Slatktine Champion, Genève, Paris, ISBN-10: 205100613X, pp: 194.

Quiniou, S., P, Cellier, T. Charnois and D. Legallois, 2013. What about sequential data mining techniques to identify linguistic patterns for stylistics? Proceedings of the 13th International Conference on Computational Linguistics and Intelligent text Processing, Mar. 11-17, CICLing, New Delhi, India, pp: 166-17. DOI: 10.1007/978-3-642-28604-9_14

Renouf, A. and J. Sinclair, 1991. Collocational frameworks in English. In: English Corpus Linguistics: Studies in the Honour of Jan Svartvik, Aijmer, K. and B. Altenberg (Eds.), Longman, London, ISBN-10: 0582059305, pp: 128-143.

Römer, U., 2009. English in academia: Does nativeness matter? Anglistik. Int. J. English Stud., 20: 89-100.

Römer, U., 2010. Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. English Text Construct., 3: 95-119. DOI: 10.1075/etc.3.1.06rom

Scott, M. and C. Tribble, 2006. Textual Patterns: Key Words and Corpus Analysis in Language Education. 1st Edn., John Benjamins, Amsterdam, ISBN-10: 9027293635, pp: 203.

Sinclair, J., 2004. Trust the Text: Language, Corpus and Discourse. 1st Edn., Routledge, London, ISBN-10: 0415317681, pp: 212.

Stubbs, M., 2007. An example of Frequent English Phraseology: Distributions, Structures and Functions. In: Corpus Linguistics 25 Years On, Facchinetti, R. (Ed.), Rodopi, Amsterdam, ISBN-10: 978904201952, pp: 89-105.

Swales, J., 1990. Genre Analysis: English in Academic and Research Settings. 1st Edn., Cambridge University Press, Cambridge, ISBN-10: 0521328131, pp: 260.

Swales, J., 2004. Research Genres: Explorations and Applications. 1st Edn., Cambridge University Press, Cambridge. ISNB-10: 0521825946, pp: 297.

Wray, A., 2002. Formulaic Language and the Lexicon. 1st Edn., Cambridge University Press, Cambridge, ISBN-10: 0521773091, pp: 332.