

# Lexical Bundles in English EU Parliamentary Discourse—Variation Across Interpreted, Translated, and Spoken Registers

Yinyin Wu

Formulas offer processing advantages in both language comprehension and production and signal one's identity in a professional community. This study attempts to explore the similarities and differences among European Parliament (EP) interpreters, translators, and Members of the EP (MEPs) in their use of English formulas, operationalized as four-word lexical bundles (LBs) identified through corpus-driven approach. English interpretations, translations, and speeches from the EP plenary sessions represent three registers, constituting the three corpora of this study. The LBs identified in each of the three corpora were then categorized based on their grammatical structures and pragmatic functions. Results showed that differences among EP interpreters, translators, and MEPs far outweigh similarities in actual bundle use, with only 10% of the total number of bundle types present in two or all three of the corpora. However, similar structural patterns of the three registers with noun phrase and prepositional phrase bundles playing a dominant role indicate informational purposes of the parliamentary discourse in general. Functional analysis revealed that the translated and spoken registers are more alike with the dominance of subject-specific bundles, indicating informational priorities. On the other hand, the interpreted register stands out with its nearly equal proportions of stance bundles, referential bundles, and subject-specific bundles, indicating a combination of informational and communicative priorities. Typology featuring LBs' functional characteristics may have pedagogical implications for translation and interpretation (T&I) training, and suggestions for enhancing LBs' pedagogical value are provided at the end of this paper.

*Keywords:* corpus-driven approach, lexical bundles, English EU parliamentary discourse, register variation, simultaneous interpreting

Received: November 3, 2020

Revised: January 11, 2021

Accepted: June 24, 2021

---

Yinyin Wu, Assistant Professor, Graduate Program in Translation and Interpretation, College of Liberal Arts, National Taiwan University, E-mail: yinyinwu@ntu.edu.tw

Ministry of Science and Technology Grant MOST 107-2410-H-002-255- funded this work. The author is grateful to the creators and distributor of "European Parliament Interpretation Corpus (EPIC), ELRA catalogue (<http://catalog.elra.info>), ISLRN: 716-168-855-843-2, ELRA ID: ELRA-S0323."

# 歐盟議會英語語篇之詞串研究—— 口譯、筆譯及演說之語域差異

吳茵茵

套語（formulas）為語言的理解與產出帶來認知處理上的優勢，亦顯示使用者所歸屬的專業社群。本研究比較歐洲議會口譯員、筆譯員與各國議員英語套語的使用差異。研究方法分為兩步驟。首先，英語套語操作化為四字詞串（lexical bundles），採用語料庫驅動（corpus-driven）的方式辨認。語料庫有三種：一、歐洲議會全體大會的英語口譯逐字稿；二、與之對應的英語筆譯文本；三、大會的英語演說逐字稿。三種語料庫分別代表三種語域（registers）。第二步驟為詞串的文法結構及語用功能分析。研究發現，在詞串的實際使用上，歐洲議會口譯員、筆譯員與各國議員的差異極大，僅 10% 的詞串類別為兩者或三者共用。然而，三種語域的文法結構呈現類似模式，皆以名詞及介系詞片語的詞串為主，顯示議會語篇（parliamentary discourse）整體而言以資訊傳達為目的。功能分析顯示筆譯與演說兩語域較為類似，皆以「主題類詞串」為主，顯示資訊功能為重。口譯語域較為特別，其「立場類詞串」、「指涉類詞串」及「主題類詞串」比例相近，顯示資訊與溝通兩種功能並重。詞串功能分類有助於口、筆譯培訓，文末討論提升詞串教學價值的建議。

關鍵詞：語料庫驅動研究方法、詞串、歐盟議會英語語篇、語域差異、同步口譯

收件：2020 年 11 月 3 日

修改：2021 年 1 月 11 日

接受：2021 年 6 月 24 日

## Introduction

Formulaic language is ubiquitous in spoken language. It is processed faster than novel language for native speakers (Conklin & Schmitt, 2012), it contributes to fluency and idiomaticity (Pawley & Syder, 1983), and it signals the speaker's identity in a given community (Wray, 2000).

Drawing inspiration from the formulaic tradition of Homeric poems, Henriksen (2007) explains that simultaneous interpreting (SI) in the European Union (EU) context is a process of oral textualization characterized by formulaic language production, which facilitates fluent and uniform interpreting output for interpreters of the same booth. To show the existence of formulas in interpreting, Henriksen recruited ten Danish staff interpreters of Joint Interpretation Service of European Commission to interpret one English speech and one German speech into Danish. It was found that although some interpreters operated with a form-based approach, providing more literal renditions, some other interpreters adopted a meaning-based approach, employing formulas for recurrent and typical ideas in the source speech. Examples of these Danish formulas include “take the lead,” “lead the way,” and “playing a role” for the following three typical and recurrent ideas in the English source speech: “put itself at the forefront,” “assuming global leadership,” and “to provide global leadership” (Henriksen, 2007, p. 10).<sup>1</sup>

Standardized structure and formulas are also used in the process of drafting legal texts and translating documents in the EU context with the help of text corpora, terminological databases, and other computer-assisted translation tools, which facilitate EU translators' work by saving time and improving consistency (Cosmai, 2007/2014; Wagner et al., 2002). Examples of these standardized

---

<sup>1</sup> These Danish formulas were translated into English by Henriksen (2007).

formulas include “after transmission of the draft legislative act to the national parliaments,” “having regard to the opinion of the Committee of the Regions,” and “acting in accordance with the ordinary legislative procedure” (Cosmai, 2007/2014, p. 118).

Formulaicity can also be observed in the EU delegates’ speeches. As Henriksen (2007) points out:

Delegates and especially Commission representatives often talk languages other than their mother tongue (primarily French or/and English), but their syntax, lexis and pronunciation will in most cases be influenced by that of their mother tongue, while their speech will include formulaic concepts specific to the EU discourse, often in French or English. The formulaic characteristics of the source text are strengthened by being produced by a non-native speaker, as such speakers tend to memorise an (at times fairly limited) number of fixed standard expressions and lack the native speaker’s command of alternatives. (p. 17)

Formulaic language seems to be a feature of interpretations, translations, and speeches in the EU context. Henriksen (2007) thus concludes, “diplomats and officials, translators and interpreters, thus all play a role in this weaving of a text of EU discourse” (p.18). The present study is an expansion and refinement of Henriksen’s study in the following three ways. First, the focus of analysis is English formulaic language produced by not only interpreters but also translators and delegates in their respective interpretations, translations, and speeches in the EU context. English has wider pedagogical implications since interpreters and translators worldwide often have English as part of their language combinations. Second, formulaic language is operationalized as lexical bundles (LBs), i.e., frequently co-occurred fixed word combinations identified objectively through corpus-driven approach. Third, the data for analysis are not elicited in an

experimental context, but speeches and simultaneous interpretations from authentic European Parliament (EP) plenary debates, as well as the translated counterparts of the simultaneous interpretations of the same source speeches. This study aims to empirically explore how EP interpreters, translators and Members of the EP (MEPs) together weave the text of English EU parliamentary discourse.

## **Corpus Studies on Lexical Bundles**

Although both fixed and semi-fixed sequences are counted as formulaic, this study focuses on frequently co-occurred fixed word sequences, such as “it goes without saying,” “the extent to which,” and “in (the) light of.” Also known as LBs, *n*-grams, or clusters in the studies of multi-word units (Greaves & Warren, 2010), these recurrent word sequences are not idiomatic in meaning, not structurally complete, not perceptually salient, but are very common and serve as building blocks in discourse (Biber & Barbieri, 2007).

Large-scale corpus-driven analyses of LBs have been conducted to describe a discourse (e.g., Jablonkai, 2010; Jalali et al., 2015), to compare disciplinary differences (e.g., Hyland, 2008b; Kashiha & Heng, 2014), or to examine register variation (e.g., Biber & Barbieri, 2007; Biber et al., 2004; Grabowski, 2015). To describe a discourse, researchers often analyze the grammatical structures and pragmatic functions of LBs identified in their respective corpora. For example, Jalali et al. (2015) examined four-word LBs in a more than two-million-word corpus of medical research articles, and found that the most common structural category is prepositional phrases signaling a time period or location, such as “in the presence of,” “at the time of,” and “in the present study,” whereas the most common functional category is text-oriented LBs, signaling transition and results and framing arguments, such as “on the other hand” and “the result of the.”

Jablonkai (2010) studied four-word LBs in a more than one-million-word corpus of English EU documents, including legal texts, legislative preparatory documents, EU funds, and institutional documents. It was found that noun phrases and prepositional phrases, such as “the entry into force” and “on behalf of the,” account for 80% of the bundle types, whereas the predominant functional category is referential bundles, which specify attributes or identify entities, such as “hereinafter referred to as,” “in the form of,” and “in accordance with the.”

Functional and sometimes structural analyses have also been used to explore disciplinary differences. Kashiha and Heng (2014) examined four-word LBs in chemistry and politics academic lectures with 25,000 words respectively. They found that both disciplines employ similar proportions of stance bundles and referential bundles, but lectures on politics rely more on discourse organizers, such as “what I’d like to” and “come up with a,” to enhance comprehensibility and achieve coherence between ideas. Hyland (2008b) compared four-word LBs in master’s theses, doctoral dissertations, and research articles with a total of 3.5 million words from four disciplines: electrical engineering, microbiology, business studies, and applied linguistics. Structural analysis revealed that noun phrase with *of*-phrase fragments comprise about a quarter of the bundle types in all four disciplines, with hard science disciplines (electrical engineering and microbiology) using more passive bundles and social science (business studies and applied linguistics) employing more prepositional phrase bundles. Functional analysis showed that hard science disciplines contain more research-oriented bundles describing research objects, contexts, equipment, materials, and environment. Examples include “the structure of the,” “in the presence of,” and “was added to the.” Social science is dominated by text-oriented bundles that frame and structure arguments, such as “in the case of,” “in the sense that,” and “will be discussed in.”

Register variation across different text types within one discipline has also been examined via comparative analyses of LBs' structures and functions. Grabowski (2015), for example, studied the 50 most frequent four-word LBs in each of the following four text types within pharmaceutical discourse with a total of more than two million words: patient information leaflets, summaries of product characteristics, clinical trial protocols, and pharmacology textbooks. It was found that the four text types differ considerably in patterns of language use due to target users, functions, and situation contexts. Patient information leaflets, having the main purpose of instructing and alerting patients, are dominated by stance bundles signaling attitudes, obligation, and desire. Summaries of product characteristics, geared towards health professionals, are dominated by referential bundles framing focus, procedure, and time. Clinical trial protocols are marked by discourse organizers, conveying purpose, design, and methodology of clinical trials. Pharmacology academic textbooks are dominated by discourse organizers and referential bundles, signaling causative-resultative links, elaborating on arguments, and referring to procedure and process. Grabowski's (2015) study shows that even within one discipline, register variation across text types exists due to differences in communicative purposes.

Biber and Barbieri (2007) examined four-word LBs extracted computationally from a wide range of university written and spoken registers. The written registers include textbooks, course management, institutional writing, and academic prose, and the spoken registers include classroom teaching, classroom management, office hours, study groups, and service encounters. Functional analysis showed that the written university registers are generally dominated by referential bundles specifying time, place, and attributes of entities; examples include "over the course of," "from the office of," and "the relationship between the." On the other hand, the spoken university registers are generally characterized by stance bundles, but only

the obligation/directive subtype (e.g., “we’re gonna have to,” “you might want to,” and “you need to take”) is prevalent in all spoken registers. Biber and Barbieri (2007) point out that functional variation of registers reflects the particular communicative needs of a given register.

These descriptive analyses of discourse, disciplinary differences, and register variation from the perspective of LBs have significant pedagogical implications for learners and teachers of a particular discipline and for translation and interpretation (T&I) training. By mastering these important yet inconspicuous building blocks in discourse, T&I trainees are able to produce spoken and written output characteristic of a particular discipline or register.

## **Research Purpose and Questions**

Interpreters and translators worldwide often need to work from and into English, which is also the case in Taiwan where most T&I practitioners and trainees are non-native English speakers. However, working into English as a B language is unavoidable especially for interpreters. The video recordings and verbatim reports of the EP plenary sessions are rich resources for studying the English output produced by professional interpreters, translators, and delegates in international conference settings. Most English booth interpreters in the EU, as well as translators working into English in the EU, are native English speakers, and therefore their renditions may have pedagogical value in the Taiwan context.

Studies based on the video recordings and/or verbatim reports of EP plenaries have been conducted to explore directionality in SI (Monti et al., 2005), the impact of SI on EU institutional hegemony (Beaton, 2007), lexical simplification (Bernardini et al., 2016), and interpreters’ visibility (Bartłomiejczyk, 2017). Although these studies have provided insights into features of the EU

parliamentary discourse, few have attempted to describe and compare EP interpretations, translations, and speeches as three registers with a focus on English formulas. Knowledge of the similarities and differences of English formulas used by interpreters, translators, and delegates allows a more targeted training for T&I novices. Interpreting trainees are encouraged to pay attention to LBs used in both interpretations and speeches to facilitate their production and comprehension of conference English, whereas translation trainees may focus on LBs used in translations and speeches, as translators sometimes have to work with scripted political and diplomatic speeches.

The present study aims to describe English EU parliamentary discourse and explore its register variation by using corpus-driven approach to identify four-word LBs in interpretations, translations, and speeches of the EP plenary sessions. The LBs identified were then analyzed structurally and functionally to reveal register variation. Both quantitative and qualitative methods were used to answer the following four questions:

1. What are the LBs shared by EP interpreters and/or translators and/or MEPs?
2. What are the structural patterns of LBs used by EP interpreters, translators, and MEPs, respectively?
3. What are the functional patterns of LBs used by EP interpreters, translators, and MEPs, respectively?
4. What are the functional characteristics of LBs used by EP interpreters, translators, and MEPs, respectively?

The purpose of research question one is pedagogically oriented. LBs commonly used by EP interpreters and/or translators and/or MEPs should be given priority in T&I training. Research questions two to four follow the convention of corpus studies on LBs. Register variation revealed via analysis of LBs' grammatical structures and pragmatic functions would inform T&I pedagogy by sensitizing trainees to language patterns specific to a given register.

## Methods

This study involved two stages. First, corpus-driven approach was adopted to identify four-word LBs crossing certain frequency thresholds. This means that the LBs were derived from data through computational tools, rather than being subjectively identified by the researcher. The second stage involved analysis of LBs' grammatical features and pragmatic functions.

### Corpus Description

English interpretations, translations, and speeches from the EP plenary sessions constitute the three corpora of the present study. Interpretation and speech corpora are from a larger corpus called European Parliament Interpretation Corpus (EPIC) (Russo et al., 2005). EPIC is a parallel corpus of EP speeches and their corresponding simultaneous interpretations in the following language combinations and directions: from English into Spanish and Italian, from Spanish into English and Italian, and from Italian into English and Spanish. This study focuses on English speeches and English interpretations only. As shown in Table 1, interpretation corpus contains 38 interpretations from Spanish or Italian into English with a total of 18,611 words, and speech corpus contains 81 English speeches with a total of 40,711 words. The transcribed speeches and simultaneous interpretations are based on the recordings of plenary sessions held on February 10th, 11th, 12th, 25th, and 26th of 2004.

It should be noted that the word counts of interpretation and speech corpora are based on pruned transcripts. The transcripts of EPIC corpus preserve all the features of spoken language, including repetitions, hesitations, fillers, and words half produced. Although these dysfluency marks are natural in interpreting and speaking, they were removed in this study for a fair comparison among interpreted, spoken, and translated corpora.

**Table 1***Constituents of the Three Corpora*

Corpus	Word count	Number of speeches	Topics
Interpretation	18,611	38	Politics (15), Justice (9), Economics and Finance (7), Health (2), Agriculture and Fisheries (2), Transport (2), Procedure and Formalities (1)
Translation	20,577	38	Same as above.
Speech	40,711	81	Politics (31), Justice (12), Economics and Finance (9), Health (23), Agriculture and Fisheries (1), Transport (1), Procedure and Formalities (4)
Total	79,899	157	

*Note.* The number of speeches for each topic is shown in parentheses.

The translation counterparts of the simultaneous interpretations of the same source speeches were copy-pasted from the final translated version of the verbatim proceedings in English published by the EP.<sup>2</sup> According to Bernardini et al. (2016), the translations of the proceedings “resulted from an independently performed translation process based on the revised verbatim reports, without any reference to the interpreters’ outputs” (p. 69). As shown in Table 1, translation corpus contains 20,577 words, about 2,000 more words than interpretation corpus. Topics on

<sup>2</sup> [https://www.europarl.europa.eu/RegData/seance\\_pleniere/compte\\_rendu/traduit/2004/02-10/P5\\_CRE\(2004\)02-10\\_DEF\\_EN.pdf](https://www.europarl.europa.eu/RegData/seance_pleniere/compte_rendu/traduit/2004/02-10/P5_CRE(2004)02-10_DEF_EN.pdf) (accessed 15 March 2019) is the online source for English verbatim proceeding dated February 10th of 2004; [https://www.europarl.europa.eu/RegData/seance\\_pleniere/compte\\_rendu/traduit/2004/02-11/P5\\_CRE\(2004\)02-11\\_DEF\\_EN.pdf](https://www.europarl.europa.eu/RegData/seance_pleniere/compte_rendu/traduit/2004/02-11/P5_CRE(2004)02-11_DEF_EN.pdf) (accessed 15 March 2019) is the online source for English verbatim proceeding dated February 11th of 2004; [https://www.europarl.europa.eu/RegData/seance\\_pleniere/compte\\_rendu/traduit/2004/02-12/P5\\_CRE\(2004\)02-12\\_DEF\\_EN.pdf](https://www.europarl.europa.eu/RegData/seance_pleniere/compte_rendu/traduit/2004/02-12/P5_CRE(2004)02-12_DEF_EN.pdf) (accessed 15 March 2019) is the online source for English verbatim proceeding dated February 12th of 2004; [https://www.europarl.europa.eu/RegData/seance\\_pleniere/compte\\_rendu/traduit/2004/02-25/P5\\_CRE\(2004\)02-25\\_DEF\\_EN.pdf](https://www.europarl.europa.eu/RegData/seance_pleniere/compte_rendu/traduit/2004/02-25/P5_CRE(2004)02-25_DEF_EN.pdf) (accessed 15 March 2019) is the online source for English verbatim proceeding dated February 25th of 2004; and [https://www.europarl.europa.eu/RegData/seance\\_pleniere/compte\\_rendu/traduit/2004/02-26/P5\\_CRE\(2004\)02-26\\_DEF\\_EN.pdf](https://www.europarl.europa.eu/RegData/seance_pleniere/compte_rendu/traduit/2004/02-26/P5_CRE(2004)02-26_DEF_EN.pdf) (accessed 15 March 2019) is the online source for English verbatim proceeding dated February 26th of 2004.

politics, justice, and economics and finance comprise the bulk for all three corpora, with speech corpus having an additional major topic on health.

## **Identification of Lexical Bundles**

Computer program WordSmith Tools version 7 (Scott, 2016) was used to identify LBs in each of the three corpora. First, LBs' length, distribution range, and frequency cut-off point had to be established. In terms of bundle length, although there have been studies exploring three- to five-word strings (e.g., Simpson-Vlach & Ellis, 2010) or three- to six-word sequences (e.g., Gilmore & Millar, 2018), most LBs studies have focused on four-word sequences (e.g., Biber & Barbieri, 2007; Biber et al., 2004; Chen & Baker, 2010; Cortes, 2004; Grabowski, 2015; Hyland, 2008a; Jablonkai, 2010; Jalali et al., 2015; Kashiha & Heng, 2014). The present study also examined four-word bundles for the following three reasons. First, aligning with previous research studies makes analysis and findings comparable. Second, three-word bundles are highly frequent whereas five-word bundles are rarer (Biber et al., 1999; Gilmore & Millar, 2018), but four-word bundles yield manageable numbers of items to work with (Chen & Baker, 2010). Third, four-word bundles provide a wider and clearer range of structures and functions to analyze than three-word bundles (Cortes, 2004; Hyland, 2008a, 2008b).

Distribution range is another criterion for identifying LBs. The purpose is to avoid idiosyncratic uses by a single author or speaker (Biber et al., 2004). To be included for analysis, LBs have been set to occur in at least three different texts or lectures (Chen & Baker, 2010; Kashiha & Heng, 2014), five different texts (Biber et al., 2004; Cortes, 2004; Jalali et al., 2015), or 10% of all texts (Hyland, 2008a, 2008b; Jablonkai, 2010) in a given corpus, or no distribution range was specified (Grabowski, 2015). Due to the sizes of the three corpora in the present study, no

distribution range was adopted as it might limit the number of LBs for subsequent analysis.

Frequency cut-off points of LBs can be arbitrary. Some studies have taken a more conservative approach by setting a high cut-off of 40 times per million words (e.g., Biber & Barbieri, 2007; Biber et al., 2004; Grabowski, 2015; Jablonkai, 2010). Some have adopted a middle position by setting the cut-off point at 20 times (Cortes, 2004; Hyland, 2008a, 2008b; Jalali et al., 2015) or 25 times (Chen & Baker, 2010) per million words. The lowest frequency cut-off point would be ten times per million words to generate a wider pool of LBs for subsequent analysis (e.g., Simpson-Vlach & Ellis, 2010).

The three corpora in the present study are very small and are of different sizes; therefore, the above-mentioned frequency cut-off points were replaced by raw frequency thresholds, as shown in Table 2. The raw frequency threshold was set at two for both interpretation and translation corpora to generate the widest pool of LBs identified for comparison. This threshold corresponds to a normalized frequency of 108 per million words in interpretation corpus, resulting in 302 bundle types with a total frequency of 704 occurrences, and 97 per million words in translation corpus, resulting in 364 bundle types with a total frequency of 918 occurrences.

Speech corpus is about twice the size of either the interpretation or translation corpus. To be more consistent with the normalized frequency of either the interpretation or translation corpus, the raw frequency threshold of speech corpus should be set at four, corresponding to a normalized frequency of 98 per million words. However, that threshold would result in only 98 bundle types, about a third of the bundle types identified in either the interpretation or translation corpus. If the same raw frequency of two (corresponding to a normalized frequency of 49 per

million words) is adopted, it would result in as many as 962 bundle types, about three times the number of bundle types in either of the other two corpora. Therefore, as shown in Table 2, a compromise was reached with a raw frequency threshold of three (corresponding to a normalized frequency of 74 per million words), resulting in 226 bundle types with a total frequency of 922 occurrences. Although the corresponding normalized frequencies of the three corpora are different, they are all above 40 per million words, the highest frequency threshold adopted in other studies.

**Table 2**

*Raw and Corresponding Normalized Frequency Thresholds of the Three Corpora and the Resulting Numbers of Bundle Types and Occurrences*

Corpus	Set raw frequency threshold	Corresponding normalized frequency (per million words)	Number of bundle types	Number of bundle occurrences
Interpretation	2	108	302	704
Translation	2	97	364	918
Speech	3	74	226	922
Total			892	2,544

## Structural and Functional Analysis of Lexical Bundles

Structural and functional analysis of LBs allows researchers to describe discourse and compare registers. The pioneers of this line of exploration are Biber et al. (1999), who categorized four-word LBs in conversation and academic prose based on bundles' grammatical structures, and a number of studies on LBs have adopted this earlier version of categorization (e.g., Chen & Baker, 2010; Cortes, 2004; Hyland, 2008a; Jalali et al., 2015). However, the structural taxonomy in

Biber et al. (2004) is more streamlined, in which four-word LBs were classified into three main types: bundles that incorporate verb phrase fragments, those that involve dependent clause fragments, and those with noun phrase and prepositional phrase fragments. This newer version of structural taxonomy was adopted by Jablonkai (2010) in her analysis of four-word LBs in English EU documents. Her added category of LBs that incorporate adjectives, adverbs, and numbers was included in the structural typology adopted in the present study, as shown in Table 3.

The functional taxonomy proposed by Biber et al. (2004) has been widely adopted by researchers (e.g., Chen & Baker, 2010; Grabowski, 2015; Jablonkai, 2010; Kashiha & Heng, 2014; Simpson-Vlach & Ellis, 2010). LBs were classified into three main categories based on their primary pragmatic functions in the corpus: Stance bundles express stance, attitudes, or assessments; discourse organizers introduce or elaborate on a topic; and referential bundles specify quantity, tangible and intangible attributes, time, space, and text markers (Biber et al., 2004).

Categorizing functions of LBs is not straightforward since a bundle may serve different functions in different contexts. Therefore, I checked the concordance lines of every bundle to verify its primary and most salient discourse-pragmatic function in the EU context. Since functions are context-based, researchers exploring different genres and registers have taken a more liberal attitude with Biber et al. (2004) taxonomy by expanding, collapsing, or modifying categories. Some of the categories added or modified by Jablonkai (2010) or by Simpson-Vlach and Ellis (2010) were included in the present study. The justification of these inclusions will be provided in descriptive analysis of LBs' functional characteristics. Table 4 presents the functional typology adopted in the present study.

**Table 3***Structural Categories of LBs*

## 1. LBs that incorporate verb phrase fragments

Subcategory	Example
1a. (connector +) 1st/2nd person pronoun + VP fragment	you don't have to
1b. (connector +) 3rd person pronoun + VP fragment	that's one of the
1c. Discourse marker + VP fragment	I mean you know
1d. Verb phrase (with non-passive verb)	take a look at
1e. Verb phrase with passive verb	can be used to
1f. Yes–no question fragment	does that make sense
1g. WH-question fragment	what does that mean

## 2. LBs that incorporate dependent clause fragments

Subcategory	Example
2a. 1st/2nd person pronoun + dependent clause fragment	I don't know why
2b. WH-clause fragment	what I want to
2c. If-clause fragment	if you have a
2d. (verb/adjective +) To-clause fragment	to come up with
2e. That-clause fragment	that I want to

## 3. LBs that incorporate noun phrase and prepositional phrase fragments

Subcategory	Example
3a. (connector +) Noun phrase with of-phrase fragment	the end of the
3b. Noun phrase with other post-modifier fragment	the way in which
3c. Other noun phrase expressions	or something like that
3d. Prepositional phrase expressions	at the same time
3e. Comparative expressions	as far as the

## 4. LBs that incorporate adjectives, adverbs, and numbers

Subcategory	Example
4a. Adjectives	the Economic and Social
4b. Adverbs	in so far as
4c. Numbers	# and # and

*Note.* Adapted from “If you look at . . . : Lexical Bundles in University Teaching and Textbooks,” by D. Biber, S. Conrad, and V. Cortes, 2004, *Applied Linguistics*, 25(3), p. 381 (<https://doi.org/10.1093/applin/25.3.371>); “English in the context of European Integration: A Corpus-Driven Analysis of Lexical Bundles in English EU Documents,” by R. Jablonkai, 2010, *English for Specific Purposes*, 29(4), p. 261 (<https://doi.org/10.1016/j.esp.2010.04.006>).

**Table 4***Functional Categories of LBs*

## I. Stance bundles

Subcategory	Example
A. Epistemic stance	the fact that the
B. Attitudinal/modality stance	
(B1) Desire	if you want to
(B2) Obligation/directive	you don't need to
(B3) Intention/prediction	I was going to
(B4) Ability/possibility	to be able to
(B5) Evaluation	it is obvious that
(B6) Hedges	it is likely that

## II. Discourse organizers

Subcategory	Example
A. Topic introduction/focus	take a look at
B. Topic elaboration/clarification	
(B1) Non-causal	you know what I'm
(B2) Cause and effect	in order to get
C. Discourse markers	thank you very much
D. Metadiscourse	go back to the

## III. Referential bundles

Subcategory	Example
A. Identification/focus	is one of the
B. Specification of attributes	
(B1) Quantity specification	a little bit of
(B2) Tangible framing attributes	the size of the
(B3) Intangible framing attributes	in the case of
C. Time/place/text-deixis/multi-functional	
(C1) Time reference	at the time of
(C2) Place reference	of the United States
(C3) Text-deixis reference	as shown in figure
(C4) Multi-functional reference	at the end of
D. Contrast and comparison	be related to the

(continued)

**Table 4***Functional Categories of LBs (continued)*

## IV. Subject-specific bundles

Subcategory	Example
A. EU-related—reference to an organization/institution	Committee of the Regions
B. EU-related—reference to a document	the Treaty on European
C. Codes	# of the EC
D. Others	the principle of subsidiarity

*Note.* Adapted from “If you look at . . . : Lexical Bundles in University Teaching and Textbooks,” by D. Biber, S. Conrad, and V. Cortes, 2004, *Applied Linguistics*, 25(3), pp. 384-388 (<https://doi.org/10.1093/applin/25.3.371>); “English in the context of European Integration: A Corpus-Driven Analysis of Lexical Bundles in English EU Documents,” by R. Jablonkai, 2010, *English for Specific Purposes*, 29(4), p. 262 (<https://doi.org/10.1016/j.esp.2010.04.006>); “An Academic Formulas List: New Methods in Phraseology Research,” by R. Simpson-Vlach and N. C. Ellis, 2010, *Applied Linguistics*, 31(4), pp. 498-502 (<https://doi.org/10.1093/applin/amp058>).

## **Lexical Bundles Shared by EP Interpreters and/or Translators and/or MEPs**

Translation corpus exhibits a wider range of different bundle types (364) than interpretation corpus does (302), as shown in Table 2. Since English translations and interpretations in this study are renditions of the same Italian or Spanish source speeches, the difference may suggest that translators, under less time pressure, can create more diversity in bundle use. This is in line with the findings of Bernardini et al. (2016) on lexical simplification of interpreted and translated EP proceedings in both Italian and English; their English interpreted and translated texts are the same as those used in the present study. They found that interpreted texts are consistently simpler than their translated counterparts, with interpreters into English making greater use of text-internal repetitions and common words.

Considering that interpretations and translations are renditions of the same source speeches, albeit one is spoken and the other is written in nature, it is surprising to see that the two corpora only share 58 bundle types, less than 1/5 of the LB types in either of the corpora (i.e., 19% of LB types in interpretation corpus and 16% in translation corpus). This suggests that the difference between EP interpreters and translators is more striking than their similarity in actual bundle use.

It is also surprising to see that interpretation and speech corpora, both are supposed to be spoken in nature, only share 35 bundle types (i.e., 12% of LB types in interpretation corpus and 15% in speech corpus). This number is close to that of the bundle types shared by translation and speech corpora—37 (i.e., 10% of LB types in translation corpus and 16% in speech corpus). It may be explained by the fact that English speeches delivered by MEPs are to some extent written in nature. In fact, 43 out of 81 speeches in speech corpus are read-out (53%), 24 are impromptu (30%), and 14 are mixed between the two (17%). The structural and functional analysis will collaborate the literate nature of EP speeches.

Table 5 is a list of 92 bundles appearing in two or all three of the corpora. Nineteen out of these 92 bundles are present in all three corpora, accounting for merely 2% of the total number of bundle types in the three corpora (892), yet the total occurrences of these 19 bundles amount to 289, accounting for 11% of the total occurrences of four-word LBs in the three corpora (2,544). If we look at the whole list, these 92 bundles account for only 10% of the total number of bundle types in the three corpora (892); however, their total occurrences (768) account for 30% of the total occurrences of LBs in the three corpora (2,544), suggesting that these 92 bundles play a somewhat important role in building the EU parliamentary discourse.

**Table 5**

*Raw Counts of Shared Bundles in Interpretation Corpus (IC), Translation Corpus (TC), and/or Speech Corpus (SC) in Total Raw Frequency Order With Bundles Present in All Three Corpora Shown in Italics*

N	Lexical bundle	IC	TC	SC	Total
1	<i>I would like to</i>	4	20	15	39
2	<i>of the European Union</i>	6	7	15	28
3	<i>the European Union and</i>	6	5	14	25
4	thank you very much	6	—	17	23
5	<i>in the European Union</i>	5	3	10	18
6	in relation to the	—	3	14	17
7	<i>on the basis of</i>	6	5	5	16
8	<i>the rule of law</i>	3	6	6	15
9	<i>that the European Union</i>	5	2	7	14
10	<i>the implementation of the</i>	2	4	8	14
11	<i>of the Lisbon strategy</i>	2	7	4	13
12	Stability and Growth Pact	—	9	4	13
13	<i>the end of the</i>	2	4	7	13
14	<i>the fight against terrorism</i>	4	6	3	13
15	<i>when it comes to</i>	4	5	4	13
16	<i>it is important to</i>	2	6	4	12
17	the area of freedom	6	6	—	12
18	<i>the Commission and the</i>	3	4	5	12
19	the Stability and Growth	—	8	4	12
20	for the European Union	—	4	7	11
21	<i>the Council and the</i>	3	2	6	11
22	as a result of	—	3	7	10
23	can only be described	5	5	—	10
24	<i>in the Member States</i>	2	4	4	10

(continued)

**Table 5**

*Raw Counts of Shared Bundles in Interpretation Corpus (IC), Translation Corpus (TC), and/or Speech Corpus (SC) in Total Raw Frequency Order With Bundles Present in All Three Corpora Shown in Italics (continued)*

N	Lexical bundle	IC	TC	SC	Total
25	only be described as	5	5	—	10
26	the work of the	2	—	8	10
27	with regard to the	—	3	7	10
28	at the same time	5	4	—	9
29	be described as terrorists	5	4	—	9
30	in the light of	5	4	—	9
31	Justice and Home Affairs	—	2	7	9
32	<i>of the European Parliament</i>	3	2	4	9
33	that there is a	2	—	7	9
34	as soon as possible	5	—	3	8
35	on the one hand	3	5	—	8
36	the European Union has	—	5	3	8
37	trafficking in human beings	2	—	6	8
38	I'd like to thank	3	—	4	7
39	in the context of	—	3	4	7
40	<i>in the fight against</i>	2	2	3	7
41	that we have to	4	—	3	7
42	to say that the	—	2	5	7
43	<i>with the European Union</i>	2	2	3	7
44	with the United States	5	2	—	7
45	I think this is	2	—	4	6
46	of the European economy	2	—	4	6
47	of the United Nations	2	4	—	6
48	on the question of	3	—	3	6

(continued)

**Table 5**

*Raw Counts of Shared Bundles in Interpretation Corpus (IC), Translation Corpus (TC), and/or Speech Corpus (SC) in Total Raw Frequency Order With Bundles Present in All Three Corpora Shown in Italics (continued)*

N	Lexical bundle	IC	TC	SC	Total
49	that we need to	2	—	4	6
50	the adoption of the	—	2	4	6
51	the European arrest warrant	2	4	—	6
52	the European Union should	2	—	4	6
53	will be able to	2	4	—	6
54	an area of freedom	—	2	3	5
55	and I'd like to	2	—	3	5
56	bear in mind that	3	2	—	5
57	before the end of	—	2	3	5
58	by the European Commission	2	3	—	5
59	human rights and democratisation	—	2	3	5
60	implementation of the Lisbon	2	—	3	5
61	implementing the Lisbon strategy	2	—	3	5
62	in all Member States	2	3	—	5
63	of the area of	3	2	—	5
64	of the Committee on	—	2	3	5
65	of the Member States	—	2	3	5
66	progress has been made	2	3	—	5
67	the context of the	—	2	3	5
68	the European Union will	—	2	3	5
69	the future of the	3	2	—	5
70	the light of the	2	3	—	5
71	the Lisbon strategy and	2	3	—	5
72	to the Council and	2	—	3	5

(continued)

**Table 5**

*Raw Counts of Shared Bundles in Interpretation Corpus (IC), Translation Corpus (TC), and/or Speech Corpus (SC) in Total Raw Frequency Order With Bundles Present in All Three Corpora Shown in Italics (continued)*

N	Lexical bundle	IC	TC	SC	Total
73	at the beginning of	2	2	—	4
74	between the European Union	2	2	—	4
75	by the Commission and	2	2	—	4
76	I think it is	2	2	—	4
77	in the case of	2	2	—	4
78	in the medium term	2	2	—	4
79	ne bis in idem	2	2	—	4
80	of the South Caucasus	2	2	—	4
81	on the guidelines for	2	2	—	4
82	rule of law and	2	2	—	4
83	say that it is	2	2	—	4
84	the Conference of Presidents	2	2	—	4
85	the fact that the	2	2	—	4
86	the President of the	2	2	—	4
87	the State and society	2	2	—	4
88	the United States and	2	2	—	4
89	this area of freedom	2	2	—	4
90	to be able to	2	2	—	4
91	we are convinced that	2	2	—	4
92	what is happening in	2	2	—	4
Total					768

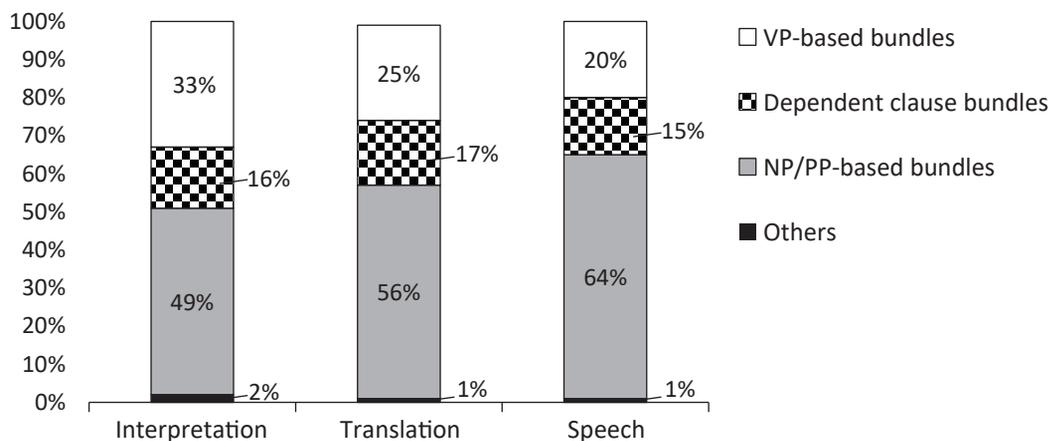
## Structural Patterns of the Three Registers

Since different normalized frequency thresholds were adopted for the three corpora, the comparison of their structural and functional patterns is discussed in terms of percentages. As Figure 1 shows, all three corpora follow similar patterns in the proportional distribution of structural categories: Noun phrase and prepositional phrase based (NP/PP-based) bundles rank as the largest category, followed by verb phrase based (VP-based) bundles, dependent clause bundles, and others, i.e., fragments containing adjectives, adverbs, or numbers. This trend of structural distribution is the same as that in Jablonkai's (2010) study on English EU documents, in which NP/PP-based bundles account for 80% of the bundle types whereas VP-based bundles almost 10% and dependent clause fragments about 6%.

According to Biber et al. (2004), NP/PP-based bundles are what set apart academic written and spoken registers from conversation, in which 90% of the bundles contain verb phrases. A register with more nouns than verbs or personal

**Figure 1**

*Structural Distribution of Lexical Bundle Types Across the Three Corpora*

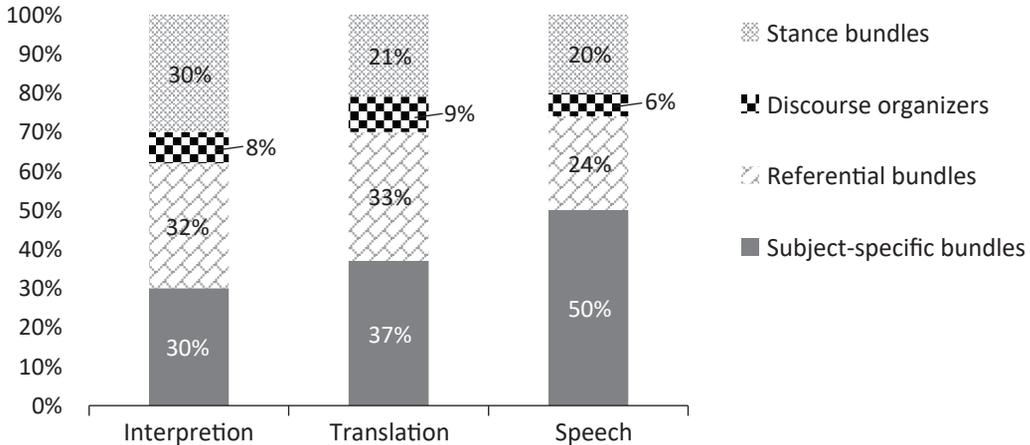


pronouns also means that the purposes are more informational than interpersonal. With NP/PP-based bundles constituting the largest structural category in all three corpora in this study, it shows that English EU parliamentary discourse is similar to English academic discourse in that both are formal English with informational purposes.

Dependent clause bundles and VP-based bundles are more common in spoken registers (conversation and classroom teaching) than in written registers (textbooks and academic prose) (Biber et al., 2004). In this study, as shown in Figure 1, dependent clause bundles account for similar proportions in all three corpora, whereas the percentage of VP-based bundles in interpretation corpus (33%) is higher than that in translation corpus (25%) and in speech corpus (20%). This indicates that the interpreted register is the most oral in nature out of the three, whereas the spoken register is the most literate with its lowest proportion of VP-based bundles and highest proportion of NP/PP-based bundles. The literate nature of the spoken register may be explained by the fact that 70% of the number of speeches in speech corpus are either read-out or a mix between read-out and impromptu (see Appendix A for raw counts and examples of bundle types across structural categories in the three corpora).

## **Functional Patterns of the Three Registers**

As Figure 2 shows, the interpreted register in the EU parliamentary discourse is unique in that stance bundles, referential bundles, and subject-specific bundles almost equally account for 1/3 of the corpus, with discourse organizers making up the remaining 8% of the bundle types. On the other hand, the translated and spoken registers display similar patterns in the proportional distribution of LB functions, with subject-specific bundles constituting the largest category (37% and 50%,

**Figure 2***Functional Distribution of Lexical Bundle Types Across the Three Corpora*

respectively), followed by referential bundles (33%, 24%), stance bundles (21%, 20%), and finally discourse organizers (9%, 6%). Both patterns (interpretation vs. translation and speech) are different from Jablonkai's (2010) analysis of EU documents, in which referential bundle types constitute the largest category (58%), followed by subject-specific bundles (33%), discourse organizers (6%), and finally stance bundles (3%).<sup>3</sup>

Referential bundles, the largest functional category in academic written registers and also extremely common in academic spoken register (classroom teaching), mainly serve informational purposes, whereas stance bundles, the largest category in spoken registers (both conversation and classroom teaching), show communicative priorities (Biber et al., 2004). If communicative and informational priorities are two ends of a continuum, English discourse of EU documents, with

<sup>3</sup> The percentages of Jablonkai's (2010) functional categories were calculated by the author of the present study based on the raw numbers of bundle types given in her Table 7 (p. 262).

stance bundles accounting for only 3% of the overall bundle types (Jablonkai, 2010), would be located at the far-right end. Similar to academic spoken register, in which both stance and referential bundles equally account for nearly 40% of the bundle types (Biber et al., 2004),<sup>4</sup> the interpreted register in the EU parliamentary discourse would be located in the middle of the continuum, combining both communicative and informational priorities. On the other hand, the translated and spoken registers in the EU parliamentary discourse are more informational than the interpreted register, but not as informational as the English of EU documents. This suggests that EP plenary debates, even in the forms of translations and scripted speeches, are still not as literate as EU documents because of the larger presence of stance bundles (see Appendix B for raw counts of bundle types across functional categories in the three corpora).

## **Functional Characteristics of Lexical Bundles in the Three Registers**

What follows is descriptive analysis of LBs' functional characteristics. LBs present in two or all three corpora are given priority in example selection.

### **Stance Bundles**

Stance bundles include two main subcategories: epistemic stance bundles and attitudinal/modality stance bundles. Epistemic stance bundles express one's knowledge about the coming information, whereas attitudinal stance bundles, containing six subtypes, frame one's attitudes.

---

<sup>4</sup> The percentages of the functional categories in Biber et al. (2004) were calculated by the author of the present study based on the raw numbers of bundle types given in their Table 4 (p. 396). It should be noted that their study does not contain the category of subject-specific bundles, which are informational in nature.

### *Epistemic Stance Bundles*

Epistemic stance bundles convey knowledge, thoughts, beliefs, viewpoints, awareness, and facts. They are mostly personal, expressing certainty (e.g., “I believe that the” and “I do not believe”) as in Example 1, or uncertainty (e.g., “I think it is,” “I’m wondering whether we,” and “as I see it”).

1. “We are convinced that” this directive, although not perfect, must be implemented during the current term of office. (translation)

Impersonal epistemic stance bundles express certainty, as in Examples 2 and 3.

2. Furthermore, we must consider “the fact that the” opening, in two years, of the Baku-Tbilisi-Ceyhan pipeline will be of enormous strategic importance. (translation)

3. In the near future, freedom, security, and justice will be a particularly critical area “taking account of the” different approaches of the new Member States on this problem. (interpretation)

As shown in Figure 3, epistemic stance bundles make up about 1/5 of the overall stance bundle types in all three corpora, suggesting their importance in both spoken and written registers.

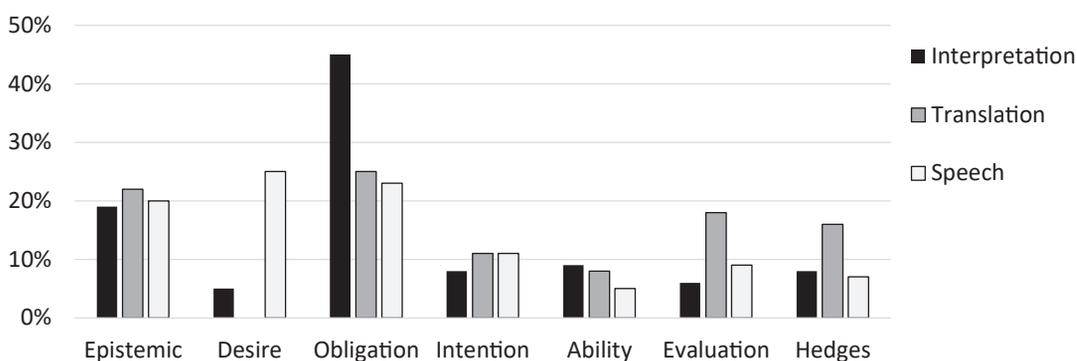
### *Attitudinal/Modality Stance Bundles*

Attitudinal/modality stance bundles express personal attitudes towards events or actions. Six subtypes were distinguished—desire, obligation/directive, intention/prediction, ability/possibility, evaluation, and hedges. The first four were created by Biber et al. (2004), and the latter two were added by Simpson-Vlach and Ellis (2010).

**Desire Bundles.** These bundles express wishes and expectations, such as “and I hope that,” “we look forward to,” and “I don’t want to.” As shown in Figure 3, desire bundles do not appear in translation corpus, whereas they account for 1/4 of

**Figure 3**

*Proportional Distribution of Stance Bundle Types Across Subcategories and Subtypes in Interpretation, Translation, and Speech Corpora*



the overall stance bundle types in speech corpus. A closer look shows that in speech corpus, desire bundles demonstrate great variety, containing words like “want” (as in “I want to thank” and “want to ensure that”), “hope” (as in “I hope that we”), “look forward to” (as in “we look forward to”), and “welcome” (as in “I very much welcome” and “I welcome the opportunity”). However, the desire bundles in interpretation corpus lack diversity, with the word “want” present in every desire bundle type (as in “I don’t want to,” “we want to do,” “we want to have,” and “we say we want”). This may suggest that MEPs giving scripted or unscripted speeches enjoy greater flexibility in expressing desire, whereas EP interpreters, under the time pressure of SI, tend to use the simplest and most direct way of expressing desire. Translation corpus containing no desire bundles is consistent with previous findings on written discourse. For example, no desire bundles are present in English EU documents (Jablonkai, 2010), and spoken classroom management contains more than triple the percentage of desire bundle types than written classroom management (Biber & Barbieri, 2007). Desire bundles may seem redundant in the eyes of EP translators, and may be deleted in the translation process.

4. “I welcome the opportunity” to brief the Members of the European Parliament today on our plans for the spring European Council. (speech)

**Obligation/Directive Bundles.** These bundles serve the functions of calling for actions (e.g., “remains to be done,” “main objective is to,” “we have to overcome,” and “to ensure that the”), drawing attention (e.g., “we must look at” and “have to look at”), or underscoring the point to be made, as in Example 5:

5. We have to “bear in mind that” citizens are kidnapped, and terrorists are in prison. (interpretation)

As shown in Figure 3, obligation bundles are particularly prominent in interpretation corpus, accounting for almost half of its stance bundle types. Interpreters use “we have to,” “we need to,” “we must,” and “we’ve got to” more than translators and MEPs do. These bundles serve the function of emphasizing the actions to be taken, but they are also less substantial words that can be used to stall for time during SI.

Obligation bundles are common in both spoken and written university registers (Biber & Barbieri, 2007), which is witnessed in the present study. In fact, as shown in Figure 3, obligation bundles also constitute the largest subtype of stance bundles in translation corpus, and the second largest in speech corpus.

Interestingly, in the analysis of Biber et al. (2004, p. 385) on university teaching, the second-person pronoun “you” is often contained in obligation bundles, directing the listener to take certain actions, such as “I want you to,” “you have to be,” “you look at the,” and “you might want to.” However, in the three EP corpora in the current study, “you” never occurs in this subtype; instead, the plural form of the first-person pronoun “we” is mostly used. This may suggest that in classroom teaching, lecturers take a more authoritative role than students, whereas MEPs, whom interpreters and translators serve, are of equal standing.

**Intention/Prediction Bundles.** These bundles express intention (e.g., “what it intends to”), agreement and support (e.g., “fully agree with the” and “support the implementation of”), or future commitment, as in Example 6:

6. We have commitments to Afghanistan, and “we will continue to” deliver on them. (interpretation)

**Ability/Possibility Bundles.** Ability bundles are the fourth subtype of attitudinal/modality bundles in Biber et al. (2004) typology, and Simpson-Vlach and Ellis (2010) added the dimension of possibility in this subtype. Ability/possibility bundles in this study introduce actual or possible actions, as in “that we can do,” “allow them to be,” and “not be possible to.” The ability bundle “be able to” in Examples 7 and 8 can be deleted without changing much of the meaning, but its presence seems to soften the coming proposition.

7. Therefore, we have “to be able to” develop ideas for the implementation of the Lisbon strategy. (interpretation)

8. I feel bound to say that the Commission ought “to be able to” put in place appropriate legal instruments to ensure easy and secure transfer of the fruits of immigrants’ labours. (translation)

**Evaluation Bundles.** Added by Simpson-Vlach and Ellis (2010), evaluation bundles express one’s judgement or attitude towards the statement to be made, as in “it is important to,” “it is essential to,” “it is incumbent on,” and “it is useful to.”

9. So it is extremely unfair and just purely demagogic “to say that the” approach of the Commission to this policy area is laxist. (speech)

**Hedges.** Also added by Simpson-Vlach and Ellis (2010), hedges express tentativeness (e.g., “have some sort of” and “appears to have been”), but more commonly, they mitigate the coming criticism by expressing one’s regrets and unwillingness (e.g., “I am sorry to” and “I have to say”) or by acknowledging the merits (e.g., “it’s all very well” and “it is true that”).

10. It's true that "progress has been made," as mentioned by the representatives of the Commission and Council and Rapporteur, and I'd like to congratulate them on that, but nonetheless the progress is inadequate. (interpretation)

## Discourse Organizers

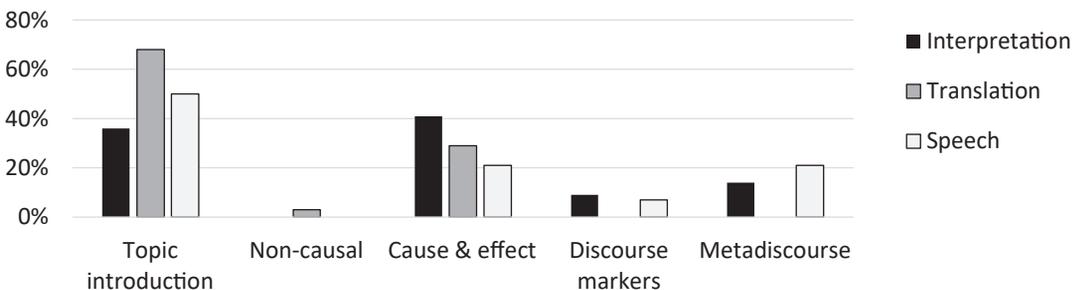
Discourse organizers include four subcategories: topic introduction/focus, topic elaboration/clarification, discourse markers, and metadiscourse. These bundles signal the start, transition, and end of a segment of discourse.

### *Topic Introduction/Focus Bundles*

Topic introduction bundles introduce a new topic. Figure 4 shows that they account for half or more than half of the bundle types under discourse organizers in both speech and translation corpora, and 1/3 in interpretation corpus. In interpretation and speech corpora, most of the topic introduction/focus bundles are extensions of "would like to," whereas those in translation corpus exhibit more diversity, such as "should like to highlight," "the third priority is," "to make a further," and "to point out that."

**Figure 4**

*Proportional Distribution of Discourse Organizing Bundle Types Across Subcategories and Subtypes in Interpretation, Translation, and Speech Corpora*



11. “I would like to” comment briefly on each of these aspects. (translation)

Example 12 illustrates the combination of two four-word bundles (“question arises as to” and “as to whether the”) into a seven-word bundle.

12. “The question arises as to whether the” figure alone is the problem.  
(translation)

### *Topic Elaboration/Clarification Bundles*

These bundles were not divided into further subtypes in Biber et al. (2004) typology, but Simpson-Vlach and Ellis (2010) created two subtypes: non-causal and cause and effect.

**Non-Causal Bundles.** Appearing only in translation corpus, these bundles clarify or paraphrase what has been mentioned previously.

13. But that protocol denies what Parliament has been requesting for years,  
“that is to say” democratic control. (translation)

**Cause and Effect Bundles.** These bundles indicate causal relationship, a reason, a purpose, or an effect. Examples include “this is why the,” “on account of the,” “in order to protect,” and “has led to a.”

14. South Africa, Argentina, Brazil, Taiwan, South Korea, for example, have all turned their backs on the proliferation of nuclear weapons partly “as a result of” international pressure, but also “as a result of” sensible and wise decisions taken in response to domestic debate and opinion. (speech)

### *Discourse Markers*

Added by Simpson-Vlach and Ellis (2010), discourse markers signal the beginning or ending of a discourse, as in “debate is now closed.” In Example 15, the bundle “thank you very much” wraps the entire speech.

15. “Thank you very much,” President. I wanna restrict my remarks to the area of legal migration. ... This is why I urge the Commissioner and the Council

to press on with this very difficult work in an important area a structured programme for planned migration for the European Union of 25 countries and more. “Thank you very much.” (speech)

### ***Metadiscourse Bundles***

Also added by Simpson-Vlach and Ellis (2010), metadiscourse is speakers/writers’ commentary on a text to interact with listeners/readers (Hyland, 2017). Bundles in this subcategory include “to my earlier answer” and “we were talking about.”

16. And here “I come back to” what Mr. Linkor has said, and also Mr. Lange also had to say. (interpretation)

Both discourse markers and metadiscourse bundles seem to be features of spoken registers, since translation corpus contains none of these two subcategories of discourse organizers, as shown in Figure 4.

### **Referential Bundles**

Drawing listeners/readers’ attention to an entity or certain attributes of an entity, referential bundles constitute the largest functional category in Jablonkai’s (2010) analysis of English EU documents. In this study, they also make up the largest functional category in interpretation corpus, and the second largest in both translation and speech corpora, as shown in Figure 2. Four main subcategories were differentiated: identification/focus, specification of attributes, time/place/text-deixis/multi-functional bundles, and contrast and comparison bundles.

#### ***Identification/Focus Bundles***

Identification/focus bundles are different from topic introduction/focus bundles under discourse organizers; the former are more local, drawing attention to a noun phrase or an idea, whereas the latter are more global, drawing attention to a

new segment of discourse. Here, focus bundles usually contain words like “it is,” “that is,” “this is,” “there is,” “there has,” “what is,” “which is,” “means that,” and “one of.” These words signal local emphasis. Focus bundles are usually followed by a noun phrase as the point of focus (e.g., “I refer to the,” “difficulty lies in the,” and “as one of the”).

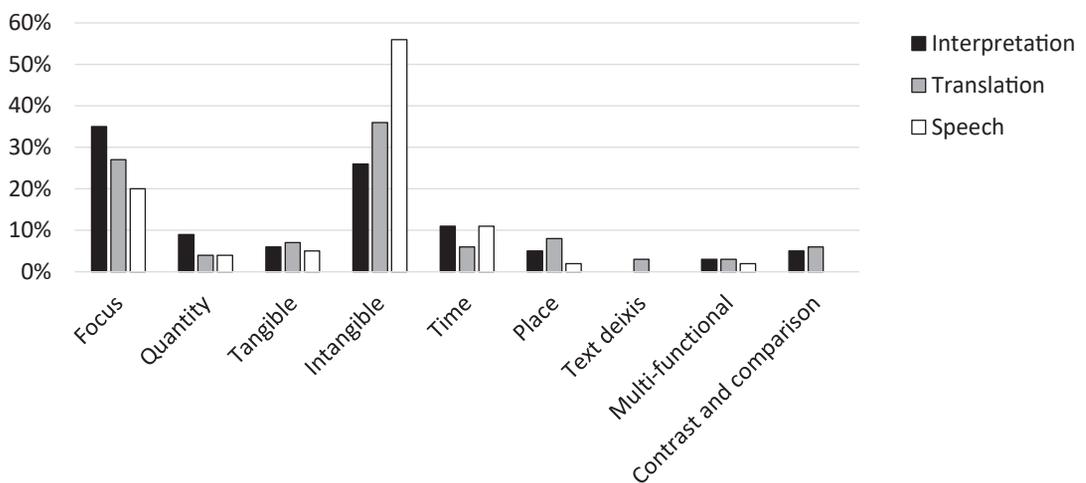
17. In order that my country may overcome this difficult time, all the European Institutions must keep a close watch over “what is happening in” Italy. (translation)

18. And what we are trying to do is to create a common level playing field at European level to guarantee that in all 25 Member States there will be the same pattern of protection of personal data “when it comes to” law enforcement. (speech)

As shown in Figure 5, focus bundles constitute the largest subcategory of referential bundles in interpretation corpus. Out of the 34 focus bundle types in interpretation corpus, eight contain fragments of simple sentence structures with

**Figure 5**

*Proportional Distribution of Referential Bundle Types Across Subcategories and Subtypes in Interpretation, Translation, and Speech Corpora*



words like “that is,” “it is,” and “this is” (as in “that’s not the way,” “it’s not just a,” and “this is something that”), whereas only two focus bundle types in either translation or speech corpus contain such words. This may again suggest that under temporal and cognitive constraints of SI, interpreters tend to resort to simpler sentence structures than translators and MEPs do.

### *Specification of Attributes*

Bundles specifying attributes have three subtypes: quantity specification, tangible framing, and intangible framing.

**Quantity Specification Bundles.** They specify quantities (e.g., “the number of those” and “majority of Member States”) or amounts (e.g., “there can be little” and “a great deal of”).

19. The Lisbon reforms must be carried through, implemented and developed “in all Member States” of the Union. (translation)

**Tangible Framing Bundles.** These bundles contain concrete words like “eye,” “face,” and “hands,” yet they may be metaphorical in meaning, as in “with an eye on,” “in the face of,” and “in the hands of.” Other tangible framing bundles refer to physical characteristics or range of the coming noun, as in “as part of the,” “in the field of,” and “this new form of.”

20. All of these subareas “of the area of” freedom, security, and justice have a direct bearing on Member States’ sovereignty. (interpretation)

**Intangible Framing Bundles.** These bundles describe abstract features, as in “the dignity of the,” “the independence of the,” and “the competitiveness of the.” They may take the form of nominalization, such as “the implementation of the,” “the development of the,” and “the creation of the.” These bundles also indicate relationships between ideas, such as “in accordance with the,” “in relation to the,” and “in connection with the.” They may also serve as bridges between clauses, as in “of which I am” and “in this way we.”

21. Without stronger institutions and better law enforcement, we'll get nowhere "in the fight against" drugs. (speech)
22. We can convince our citizens that this needs to be done to improve security "on the basis of" more cooperation and not "on the basis of" more repression. (interpretation)

Intangible framing bundles constitute the largest subtype under referential bundles in Jablonkai's (2010) analysis of EU documents. In this study, intangible framing bundles also make up the largest subtype of referential bundles in both translation and speech corpora, and the second largest subtype in interpretation corpus, as shown in Figure 5. This may suggest that intangible framing bundles are common in formal English, both written and spoken.

### *Time/Place/Text-Deixis/Multi-Functional Bundles*

**Time Bundles.** They refer to a specific time (e.g., "February of this year"), a period of time (e.g., "over the next three" and "in the last few"), the onset or ending of time (e.g., "from this point of" and "before the end of"), or immediacy (e.g., "at the earliest opportunity" and "as soon as possible").

23. And I'm sure "in the medium term," if we have that, we can get to budgetary balance "in the medium term," especially if we look at the population trends. (interpretation)

**Place Bundles.** They refer to countries or regions.

24. Mr. President, the problems "of the South Caucasus" are as easy to identify as they are difficult to solve. (translation)

**Text-Deixis Bundles.** These bundles are unique to translation corpus, as shown in Figure 5. This is consistent with the findings of Biber et al. (2004) in that text-deixis bundles are common only in written registers.

25. "As stated in the" report, the pillar structure should be done away with. (translation)

**Multi-Functional Bundles.** Depending on the context, multi-functional bundles may refer to time or concepts, act as discourse markers, or be part of an idiom (e.g., “at the end of” the day). Bundle “at the same time” acts as a non-causal topic elaboration bundle (the same function as “furthermore”) in Example 26, and acts as a time referent in Example 27.

26. But the current Pact has rules that are far too rigid to enable us to do any of that, so we have to rethink it. “At the same time,” it must provide a guarantee of stability in Europe, which will be safeguarded by the carefully and conscious choices taken by our leaders. (interpretation)

27. Its [the draft law’s] development is happening “at the same time” as the action on the vigilante groups. (interpretation)

### *Contrast and Comparison Bundles*

These bundles signal similarities or differences between two entities (e.g., “the same is true,” “is more constructive than,” and “a comparison between the”), or highlight two different opinions about something, as in the case of “on the one hand” and “on the other hand.” Both interpretation and translation corpora have the bundle “on the one hand,” which interestingly is not followed by “on the other hand” in interpretation corpus. It is possible that under the time constraint of SI, interpreters either forget that they have started the argument with “on the one hand,” or they fall back on shorter words like “but” to save time, as in Example 28:

28. I’m very sorry about the ambiguity and incoherence being expressed by some colleagues, who “on the one hand” praised the report because it’s good, but then very illogically say that they are not going to support it. (interpretation)

Example 29 is the translation counterpart.

29. I very much regret the seriously misguided and inconsistent way in which certain Members, “on the one hand,” heap great praise on the report because it is good, but, “on the other,” say that they are not going to support it. (translation)

## Subject-Specific Bundles

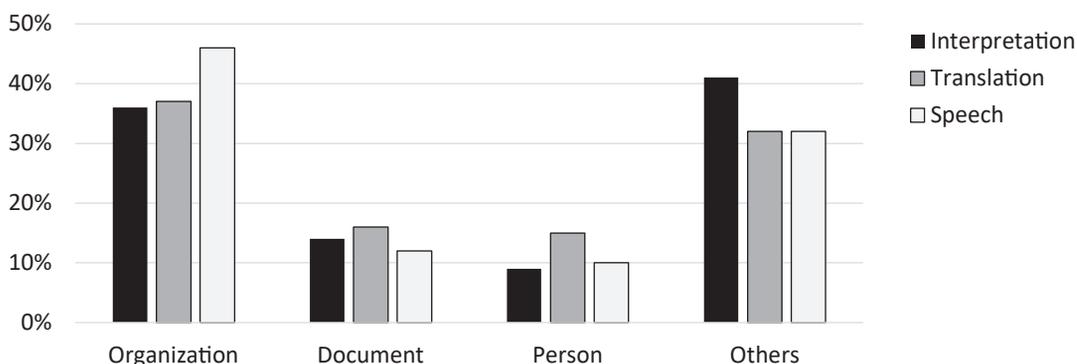
Created by Jablonkai (2010), the category of subject-specific bundles is devoted solely to EU-related terminology or concepts. Four subcategories were differentiated: reference to an organization/institution, reference to a document, reference to a person, and others. As shown in Figure 6, reference to an organization and others are the two major subcategories in all three corpora.

### *EU-Related—Reference to an Organization/Institution*

Jablonkai (2010) included reference to a country in this subcategory, but it overlaps with place bundles under referential bundles. In the present study, reference to a country was categorized as a place bundle, whereas the bundles in this subcategory refer to EU institutions (e.g., “the Commission and Council” and

**Figure 6**

*Proportional Distribution of Subject-Specific Bundle Types Across Subcategories in Interpretation, Translation, and Speech Corpora*



“the European Parliament in”), international organizations (e.g., “of the United Nations”), European governments (e.g., “that the UK government”), and Europe as a whole (e.g., “of the enlarged Europe”).

30. I’d like to thank you for this opportunity to discuss the work of the “Justice and Home Affairs” Council in 2003. (speech)

### ***EU-Related—Reference to a Document***

These bundles refer to treaties, initiatives, charters, reports, agreements, strategies, guidelines, and pacts.

31. We must build a dynamic in the Lisbon process by ensuring that the “Stability and Growth Pact” actually provides for growth as well as stability. (speech)

### ***EU-Related—Reference to a Person***

This subcategory was created for the present study, replacing Codes in Jablonkai’s (2010) typology. Bundles referring to a person contain words such as “members,” “ministers,” “President-in-Office,” “Vice-President,” “representatives,” “rapporteur,” “researchers,” “commissioner,” “citizens,” and “gentleman.”

32. That will be for the new Commission to decide on, and particularly “the President of the” Commission. (interpretation)

### ***Others***

These bundles express specific EU-related issues or concepts, usually in the form of noun phrases. Examples include “cooperation among the police,” “the cease of hostilities,” and “democracy and human rights.”

33. Our success will depend on standing firm in respecting “the rule of law” and applying all the mechanisms permitted by law. (interpretation)

## Conclusions

This study investigated the similarities and differences across interpreted, translated, and spoken registers in English EU parliamentary discourse by identifying frequently occurred four-word LBs through corpus-driven approach and by analyzing LBs' grammatical structures and pragmatic functions. It was found that only around 10% of the bundle types are shared among EP interpreters and/or translators and/or MEPs, suggesting differences far outweigh similarities in actual bundle use.

However, the three registers display similar patterns in structural distribution. The dominance of NP/PP-based bundle types in all three corpora suggests that the English discourse produced by EP interpreters, translators, and MEPs is highly informational. What is surprising is that speeches delivered by MEPs are more literate in nature than not just interpretations but also translations, as shown by the highest proportion of NP/PP-based bundle types and the lowest proportion of VP-based bundle types in speech corpus. The literate nature of the spoken register may be explained by the fact that 70% of the speeches in speech corpus are either read-out or a mix between read-out and impromptu.

Functional analysis revealed that the interpreted register contains almost equal proportions of subject-specific bundles, referential bundles, and stance bundles, indicating a combination of informational and communicative priorities. On the other hand, the spoken and translated registers display similar functional patterns. The more dominant role played by subject-specific bundles and referential bundles indicates informational priorities of the two registers. Although the spoken register seems to be more literate than even the translated register, as shown by structural analysis, a closer examination of bundles' functional characteristics revealed that the spoken register still possesses oral features with its high proportion of desire

bundle types, the presence of discourse markers and metadiscourse bundles, and the absence of text-deixis bundles.

The four-word bundles identified in the present study might provide practical value for citizens, translators, and interpreters of the EU Member States. These bundles may also be useful for translators and interpreters worldwide whose language combinations involve English, as prefabricated sequences signal one's professional identity and offer processing advantages in both language production and comprehension.

To enhance bundles' pedagogical value, two major limitations of the present study may need to be addressed. First, compared with the corpus sizes of other corpus-driven studies, the sizes of the three corpora in this study are very small. Although small corpora allow close examination of patterns of language use in context (Koester, 2010), language patterns revealed by large corpora warrant further study. Second, the number and frequency of four-word bundles identified in the present study may be inflated, as bundles may overlap. For example, in interpretation corpus, the bundle "in the fight against" occurs twice, whereas the bundle "the fight against terrorism" occurs four times, amounting to a raw frequency of six. However, the two bundles can be collapsed as one five-word bundle, "(in) the fight against terrorism," with a frequency count of four. But this five-word bundle would render comparison of LBs across the three corpora difficult. For example, in speech corpus, only the bundle "in the fight against" makes the frequency cut, and in each occurrence, it is followed by a different object: "drugs," "terrorism," or "international organised crime and terrorism." For pedagogical purposes, collapsed bundles with variable slots, such as "(in) the fight against terrorism/drugs/international organised crime," may be more complete in meaning and thus easier for T&I trainees to memorize. Future studies are

encouraged to collapse bundles before functional categorization for pedagogical applications.

Two other directions also merit further investigation. First, other features of English EU Parliamentary discourse, such as lexical variety, lexical density, collocations, and sentence length, may be explored via corpus-based or corpus-driven approach to further reveal the universality and variation of the three modalities: interpretation, translation, and speech. Second, the interpreted, translated and spoken registers of other languages, such as Mandarin Chinese, may be examined using corpus-based or corpus-driven approach to understand intra-language and inter-language register variation for the purposes of T&I training.

## References

- Bartłomiejczyk, M. (2017). The interpreter's visibility in the European Parliament. *Interpreting*, 19(2), 159-185. <https://doi.org/10.1075/intp.19.2.01bar>
- Beaton, M. (2007). Interpreted ideologies in institutional discourse: The case of the European Parliament. *The Translator*, 13(2), 271-296. <https://doi.org/10.1080/13556509.2007.10799241>
- Bernardini, S., Ferraresi, A., & Miličević, M. (2016). From EPIC to EPTIC — Exploring simplification in interpreting and translation from an intermodal perspective. *Target*, 28(1), 61-86. <https://doi.org/10.1075/TARGET.28.1.03BER>
- Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26(3), 263-286. <https://doi.org/10.1016/j.esp.2006.08.003>
- Biber, D., Conrad, S., & Cortes, V. (2004). *If you look at . . .*: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371-405. <https://doi.org/10.1093/applin/25.3.371>
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman.
- Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, 14(2), 30-49. <http://doi.org/10.125/44213>
- Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, 32, 45-61. <https://doi.org/10.1017/S0267190512000074>
- Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23(4), 397-423. <https://doi.org/10.1016/j.esp.2003.12.001>

- Cosmai, D. (2014). *The language of Europe: Multilingualism and translation in the EU institutions: Practice, problems and perspectives* (D. A. Best, Trans.; 3rd ed.). Editions de l'Université de Bruxelles. (Original work published 2007)
- Gilmore, A., & Millar, N. (2018). The language of civil engineering research articles: A corpus-based approach. *English for Specific Purposes, 51*, 1-17. <https://doi.org/10.1016/j.esp.2018.02.002>
- Grabowski, Ł. (2015). Keywords and lexical bundles within English pharmaceutical discourse: A corpus-driven description. *English for Specific Purposes, 38*, 23-33. <https://doi.org/10.1016/j.esp.2014.10.004>
- Greaves, C., & Warren, M. (2010). What can a corpus tell us about multi-word units? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 212-226). Routledge.
- Henriksen, L. (2007). The song in the booth: Formulaic interpreting and oral textualisation. *Interpreting, 9*(1), 1-20. <https://doi.org/10.1075/intp.9.1.02hen>
- Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41-62. <https://doi.org/10.1111/j.1473-4192.2008.00178.x>
- Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4-21. <https://doi.org/10.1016/j.esp.2007.06.001>
- Hyland, K. (2017). Metadiscourse: What is it and where is it going? *Journal of Pragmatics, 113*, 16-29. <https://doi.org/10.1016/j.pragma.2017.03.007>
- Jablonkai, R. (2010). English in the context of European integration: A corpus-driven analysis of lexical bundles in English EU documents. *English for Specific Purposes, 29*(4), 253-267. <https://doi.org/10.1016/j.esp.2010.04.006>

- Jalali, Z. S., Moini, M. R., & Arani, M. A. (2015). Structural and functional analysis of lexical bundles in medical research articles: A corpus-based study. *International Journal of Information Science and Management*, 13(1), 51-69. <https://ijism.ricest.ac.ir/index.php/ijism/article/view/503>
- Kashiha, H., & Heng, C. S. (2014). Discourse functions of formulaic sequences in academic speech across two disciplines. *GEMA Online® Journal of Language Studies*, 14(2), 15-27. <http://doi.org/10.17576/GEMA-2014-1402-02>
- Koester, A. (2010). Building small specialised corpora. In A. O’Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 66-79). Routledge.
- Monti, C., Bendazzoli, C., Sandrelli, A., & Russo, M. (2005). Studying directionality in simultaneous interpreting through an electronic corpus: EPIC (European Parliament Interpreting Corpus). *Meta*, 50(4). <https://doi.org/10.7202/019850ar>
- Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191-226). Addison Wesley Longman.
- Russo, M., Bendazzoli, C., Monti, C., Sandrelli, A., Baroni, M., Bernardini, S., Mack, G., Piccioni, L., Zanchetta, E., Ballardini, E., & Mead, P. (2005). *European Parliament interpretation corpus (EPIC)* (Version 1) [Data set]. The European Language Resources Association. <https://catalog.elra.info/en-us/repository/browse/ELRA-S0323/>
- Scott, M. (2016). *WordSmith tools* (Version 7) [Computer software]. Lexical Analysis Software. <https://www.lexically.net/wordsmith/>
- Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31(4), 487-512. <https://doi.org/10.1093/applin/amp058>

- Wagner, E., Bech, S., & Martínez, J. M. (2002). *Translating for the European Union institutions*. St. Jerome.
- Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics*, 21(4), 463-489. <https://doi.org/10.1093/applin/21.4.463>

## Appendix A

*Raw Counts and Examples of Bundle Types Across Structural Categories in Interpretation Corpus (IC), Translation Corpus (TC), and Speech Corpus (SC)*

Category	Subcategory	IC	TC	SC	Example
1. VP-based	1a. 1st/2nd person pronoun + VP fragment	44	19	28	we have to overcome
	1b. 3rd person pronoun + VP fragment	25	42	8	debate is now closed
	1d. Verb phrase with non-passive verb	22	18	7	taking account of the
	1e. Verb phrase with passive verb	7	13	—	be borne in mind
	1f. Yes–no question fragment	—	—	2	are there any comments
	1g. WH-question fragment	1	—	—	how do we get
	Sub-total		99	92	45
2. Dependent clause	2b. WH-clause fragment	6	9	2	what is happening in
	2c. If-clause fragment	3	—	1	if we do that
	2d. To-clause fragment	23	42	13	to face up to
	2e. That-clause fragment	17	11	17	bear in mind that
	Sub-total		49	62	33
3. NP/PP-based	3a. Noun phrase with of-phrase fragment	46	56	28	the cease of hostilities
	3b. Noun phrase with other post- modifier fragment	18	32	22	our policies with the
	3c. Other noun phrase expressions	23	28	31	the State and society
	3d. Prepositional phrase expressions	59	88	65	as part of the
	3e. Comparative expressions	2	1	—	than in the United
	Sub-total		148	205	146
4. Others	4a. Adjectives	3	3	—	aware of the need
	4b. Adverbs	2	2	2	here in this House
	4c. Numbers	1	—	—	half a per cent
	Sub-total		6	5	2
Total		302	364	226	

## Appendix B

*Raw Counts of Bundles Types Across Functional Categories in Interpretation Corpus (IC), Translation Corpus (TC), and Speech Corpus (SC)*

Category	Subcategory	IC	TC	SC
I. Stance expressions	A. Epistemic stance	17	17	9
	B. Attitudinal/modality stance			
	(B1) Desire	5	—	11
	(B2) Obligation/directive	41	19	10
	(B3) Intention/prediction	7	8	5
	(B4) Ability/possibility	8	6	2
	(B5) Evaluation	6	14	4
	(B6) Hedges	7	12	3
Sub-total		91	76	44
II. Discourse organizers	A. Topic introduction/focus	8	23	7
	B. Topic elaboration/clarification			
	(B1) Non-causal	—	1	—
	(B2) Cause and effect	9	10	3
	C. Discourse markers	2	—	3
	D. Metadiscourse	3	—	1
Sub-total		22	34	14
III. Referential expressions	A. Identification/focus	34	32	11
	B. Specification of attributes			
	(B1) Quantity specification	9	5	2
	(B2) Tangible framing attributes	6	8	3
	(B3) Intangible framing attributes	25	43	30
	C. Time/place/text-deixis/multi-functional			
	(C1) Time reference	10	7	6
	(C2) Place reference	5	9	1
	(C3) Text-deixis reference	—	3	—
	(C4) Multi-functional reference	3	4	1
	D. Contrast and comparison	5	7	—
Sub-total		97	118	54

(continued)

*Raw Counts of Bundles Types Across Functional Categories in Interpretation Corpus (IC), Translation Corpus (TC), and Speech Corpus (SC) (continued)*

Category	Subcategory	IC	TC	SC
IV. Subject-specific bundles	A. EU-related—reference to an organization/ institution	33	50	53
	B. EU-related—reference to a document	13	22	14
	C. EU-related—reference to a person	8	20	11
	D. Others	38	44	36
Sub-total		92	136	114
Total		302	364	226