transLogos 2024 Vol 7 Issue 1
Lee, Tzu-yi Elaine, pp. 41–61
https://dx.doi.org/10.29228/transLogos.64

Exploring Naturalness in MOOC Video Lecture MT Subtitles: A Case Study

Tzu-yi Elaine LEE*

This case study investigates the naturalness of subtitle translation using massive open online course (MOOC) video lecture subtitles that were machine translation (MT)-generated by a computer-assisted translation (CAT) tool, *Termsoup*, after the pre- and post-editing phases. Both the pre- and post-editing phases were conducted by translator trainees while supervised by a professional translation trainer. With the accuracy secured, the naturalness of the MOOC translated subtitles from Chinese to English is further explored. In this study, distributions of lexical bundles (LBs) in terms of structural and functional properties are used to investigate the academic register in the MT output. The result shows that while in previous literature classroom teaching extensively uses a taxonomy of bundles in both written and spoken modes, the investigated corpus tends toward the 'literate' side rather than the 'oral' side. Nonetheless, the increase in the number of discourse organizers discovered in the corpus not only implies that the MOOC language may share comparable purposes with TED talks but also aids in the formation of spoken elements in the register. The study ends with suggestions for extending the corpus of MOOCs to gain a better grasp of specific formulaic languages in subtitle translation, along with proposals for pre-editing the source text to enhance the academic register before the MT process.
Keywords: naturalness of subtitles; massive open online course video lectures; lexical bundles; academic register; machine translation output

1. Introduction

With the rapid advancement in new media in recent years, audiovisual translation (AVT) has grown in popularity and accessibility for a wide range of purposes that go beyond traditional media and AVT contexts. AVT is progressively being integrated with language technologies such as computer-assisted translation (CAT) tools, machine translation (MT), or automated subtitling. Indeed, the increasing exposure of AVT in language technology is evident in the natural language processing field. While the field of AVT has grown significantly in the last two decades, with a recent emphasis on empirical research and reception studies (e.g., Doherty

* Associate professor at Chung Yuan Christian University, Taoyuan.
E-mail: t.y.lee@cycu.edu.tw; ORCID ID: http://orcid.org/0000-0002-1139-9294.
(Received 18 April 2024; accepted 14 June 2024)

and Kruger 2018; Perego 2016), the interaction of AVT and language technology has received less attention, owing to a lack of interaction between the two fields of research.

Early research with MT in subtitle workflows focused on creating systems to generate MT subtitles for broadcast (Georgakopoulou 2019). Other similar systems, such as rules-based machine translation (RBMT) technology or example-based MT, were developed later to translate closed captions. Furthermore, several European research projects, such as transLectures and TraMOOC, have attempted to address the issue of MT in offline subtitle workflows, with the goal of serving the needs of millions of students worldwide through MT for academic video lectures in the form of subtitles. Sharon O'Brien (2017) posits that human engagement with MT involves evaluation, correction, and use. Evaluation and revision involve pre- and post-editing (Arenas 2019) as intermediate processing phases, whereas use refers to end-user processing.

Massive open online courses (MOOCs) are audiovisual assets that have revolutionized higher education teaching and learning around the world. Open to learners worldwide, MOOCs have become one of the most convenient platforms offering high-quality learning resources (Uchidiuno et al. 2018) and interactive learning experiences (Wang et al. 2019). The increasing popularity of MOOCs offers opportunities for learners to obtain new skills and abilities without time and space restrictions (Vorobyeva 2018). Some universities in Taiwan have started establishing their own MOOC platforms to offer online courses, like MIT and Stanford, while some others collaborate with other universities to offer online courses and micro-credentials. According to Xiaoli Yu (2022), few studies have explored the language used in MOOC lecture videos and instructors' discourse from a linguistic perspective (e.g., Atapattu and Falkner 2018; Das and Das 2019; Shi et al. 2019, 11–19), let alone these English-medium MOOC video lectures from the perspective of AVT and MT.

Lectures are among the most influential elements in the intellectual growth of university students. Studies on university language have indicated that informational communication is "performed via very diverse linguistic strategies in speech and writing" (Biber 2006b, 222). Moreover, much scholarship has discussed extensive linguistic features and language variation in university lectures (e.g., Biber 2006a, 2006b; Biber, Conrad, and Cortes 2004; Crawford Camiciottoli 2007; Csomay 2006; Nesi and Basturkmen 2006; Lee and Subtirelu 2015). According to previous research, English-medium university lectures in various contexts are

delivered under time constraints (Csomay 2006), share similar linguistic features with conversations and academic writing (Biber 2006a, 2006b; Csomay 2006), and exhibit cross-disciplinary differences in language use when presenting instructional materials (Csomay 2007).

In contrast, free from time constraints, video lectures such as MOOCs are mostly pre-recorded, ranging from five to 20 minutes. This relatively short duration differentiates the concept of MOOC video lectures from that of traditional university lectures. Nevertheless, the videos display academic content in an academic context and, thus, are still considered lectures (Yu 2022). Video lectures are typically accompanied by visual slides and other reading materials. Despite differences between MOOC video lectures and traditional university lecturers in terms of the participants, relationships among participants, channel, processing circumstances, and setting (Biber and Conrad 2019), both are delivered in spoken mode and designed to verbally present the main learning points; thus, there should be similarities in terms of naturalness for both registers.

Meanwhile, as online education such as MOOCs have become more extensively diffused and recognized, particularly considering the impact of the worldwide pandemic in 2020, the instructional language of MOOC video lectures has not been widely examined, although some scholars (e.g., Atapattu and Falkner 2018) are starting to highlight interaction peaks on account of certain discourse features, along with other studies on non-verbal instructional features (Sharma et al. 2016), quality aspects of videos (Shi et al. 2019), or video designing issues (Wang, Chiu, and Lee 2020). Nevertheless, there has still been little research in the field of AVT investigating the newly developing academic spoken register after MT, let alone its naturalness.

The present study borrows the idea of John Sinclair (1995, 833) that "naturalness" is utilized as a synonym for "idiomaticity" to represent the nativelike choice of a phrase in a particular situation. Therefore, this case study explores the naturalness of the academic spoken register from a corpus of MOOC subtitles translated from Chinese to English as the MT output via lexical bundles (LBs) (Biber, Conrad, and Cortes 2004), which will be introduced in the next section.

2. Naturalness in AVT and LBs

Little research was carried out on the topic of AVT before the start of the new millennium, when the situation underwent a radical turn and publications in interlingual translation, as well as accessibility, started to emerge (Días-Cintas 2019; Bogucku 2019; Kuo 2020). Moreover, some scholars have started to propose models (e.g., Chiaro 2008; Romero-Fresco 2011; Pedersen 2017) or templates (Nikolic 2015; Chaume-Varela 2007) to gauge quality from theoretical and practical perspectives. Unlike standard Translation Quality Assessment (TQA) methods and metrics, AVT evaluation is mostly dependent on prescriptive industry rules that vary by company, medium, location, language, and nation (Doherty and Kruger 2018). Nevertheless, commonalities can be discovered in the broad dimensions of accuracy, presentation, and timing. Of course, the linguistic information at the heart of the audiovisual text is a critical component of this assessment in pursuit of quality and accuracy in the MT of audiovisual content, which may include pre-editing and post-editing (Arenas 2019; Doherty 2017).

In addition to accuracy, previous research has demonstrated variable degrees of congruence between translated audiovisual speech and real conversation, with linguistic elements serving as preferred bearers of orality in dubbing (Pavesi 2008, 2009, 2013; Pavesi, Formentelli, and Ghia 2014; Pettit 2005; Romero-Fresco 2009; Valdéon 2008). While few studies have been conducted on the naturalness of subtitles or the correspondence of registers (e.g., Pettit 2005), the current study was inspired by these studies. The premise is that language is best analyzed by comparing it to the register it imitates if its peculiar features are considered. In this way, the MT output of MOOC subtitles explored in this case study should, to some extent, resemble the real-life university-level academic register to evoke reality (Romero-Fresco 2009), as regarded by Luis Pérez-González as “real realism” (2007, 7). This real realism is to be achieved through natural and spontaneous-sounding lectures to facilitate the comprehension of learners.

The current study uses LBs, or chunks in general, as the unit of analysis to explore the naturalness of subtitles. LBs are defined by Biber, Conrad, and Leech (2002, 190) as “recurrent expressions, regardless of their idiomaticity or structural status.” In other words, LBs are sequences of word forms that often occur in natural language and could be comparable to those of “chunks,” “sequences,” “formulaic language,” or “multi-word units” (Ellis, Simpson-Vlach,

and Maynard 2008; Wood 2010; Wray and Perkins 2000). LBs are generally characterized by three unique features, with the first being their high frequency. Word combinations must often occur in a specific register or corpus to qualify as an LB. Second, LBs provide significant online processing benefits by boosting predictability within a certain context and simplifying understanding and output in both first- and second-language use (Arnon and Snider 2010; Ellis Simpson-Vlach, and Maynard 2008; Ellis and Simpson-Vlach 2009; Li and Schmitt 2009; Tremblay et al. 2011).

Finally, LBs have been demonstrated to be "text building blocks" (Biber, Conrad, and Leech 2002, 443), which can take on three structural forms (Biber, Conrad, and Cortes 2004, 380): verb phrase fragments (VP), dependent clause fragments (DC), and nominal or prepositional phrase components (NP). Although most LBs are not as semantically or structurally comprehensive as idioms or collocations, they either begin or finish at a sentence or phrase boundary and serve as "a kind of discourse anchor," which tells the listeners or students how to interpret the information regarding stance, discourse organization, or referential status (Biber, Conrad, and Cortes 2004, 377–399).

LBs are common in academia, although their application differs among fields and registers, such as classroom instruction, class discussion, and writing tasks (e.g., Biber, Conrad, and Cortes 2004; Biber and Barbieri 2007; Cortes 2004; Hyland 2008; Salazar 2011; Wang 2017). Some academics have studied the many discourse functions of LBs to better understand how speakers or writers build links, define traits, make judgments, and structure discourse in academic registers (e.g., Biber, Conrad, and Cortes 2004; Coxhead, Dang, and Mukai 2017). According to Chen-Yu Liu and Hao-Jan Howard Chen (2020a, 2020b), an increasing number of researchers (e.g., Simpson 2004; Neely and Cortes 2009; Nesi and Basturkmen 2006) have utilized corpora of spoken and written academic language to explore the use of LBs in academic spoken registers.

Meanwhile, despite the presence of numerous bundle studies on academic registers, few have investigated the variance in LBs among spoken academic registers (Biber, Conrad, and Cortes 2004; Biber and Barbieri 2007; Liu and Chen 2020a, 2020b; Nesi and Basturkmen 2006). Douglas Biber, Susan Conrad, and Viviana Cortes (2004) conducted a notable study comparing the bundles used in classroom teaching, textbooks, academic prose, and conversation. They discovered that classroom instruction depends on LBs linked with both spoken and written registers and makes extensive use of LBs that express declarative and interrogative phrase

fragments. Moreover, like academic writing and textbooks, it makes extensive use of noun phrase and prepositional phrase LBs. Liu and Chen (2020b) compared functional variants of LBs in academic lectures and TED speeches, in addition to examining the purposes of LBs in undergraduate academic lectures for instructional application. The findings most relevant to the present study are that LBs most often perform referential and stance functions in academic lectures, while their function as discourse organizers is also important. Moreover, in comparison to TED talks, academic lectures use more stance bundles to show their intention and to inform the audience about upcoming events. The taxonomy will be fully introduced in the following section.

Due to the feature of formulaicity, LBs have recently been gaining attention in translation and interpretation research (Aston 2018; Li and Halverson 2020, 2022; Plevoets and Defrancq 2018), but they have rarely been considered in the AVT context. Other applications can be found in the studies by Kanglong Liu, Joyce Oiwun Cheung, and Riccardo Moratto (2022) and Kanglong Liu and Muhammad Afzaal (2021), who used LBs as an indicator to explore two translators' styles in English translations of fictional dialogues in *Hongloumeng* via a corpus-assisted approach. Changsoo Lee (2013) similarly reported that the use of LBs as a probing methodology, albeit different from the standard method, can be critical in exploring diversity in stylistic choices between translators. In addition, LBs as building blocks are relevant to legal language and legal translation (e.g., Berūkštienė 2017; Biel 2018). More importantly, Łukasz Grabowski (2018) investigated the use of bilingual LBs to improve the degree of naturalness and textual fit of translated texts in a purpose-designed comparable corpus of English and Polish patient information leaflets by identifying recurrent sequences of 3-7 grams in functional types. The findings imply that bilingual LBs generated from similar corpora have untapped potential for MT, CAT, and bilingual lexicography.

Whereas the aforementioned studies using LBs employed a frequency-driven approach, the present study opts for a qualitative approach due to a relatively smaller corpus (MOOC subtitles translated into English via MT). As Biber (1988) points out, a qualitative approach is also necessary to reach specification, avoid contradictions, and supplement quantitative analysis in previous studies. Accordingly, this study addresses the following questions: (i) How do LBs distributed in MOOC translated subtitles affect the naturalness of academic register? (ii) How is the academic register of MOOC subtitles preserved in MT translated output? Is there any potential solution for modification, if necessary?

3. Corpus and Methodology

The MOOC applied in this case study is entitled Introduction to Medical Devices and Principles, currently taught by a professor from the Department of Biomedical Engineering at a university in northern Taiwan. The original language used by the lecturer is Chinese, and the MOOC is about six hours long, split into 31 approximately 15-minute videos. The Chinese subtitles are transcribed into Word documents totaling 98,939 words (Chinese characters) and 81,536 words in English after MT. The MOOCs platform established at this Taiwanese university is in its early stages, and the school encourages professors from different departments to present their courses as videos for both local and international students, particularly to address issues related to the COVID-19 pandemic. As the main language for the MOOCs is mostly Chinese, most of these MOOCs are currently in the translation (mostly by humans) process, but the course selected for investigation in the present study is one of the few using MT available. Moreover, the MOOC subtitles were translated by trainee translators via a CAT tool, *Termsoup*, and went through pre-editing and post-editing processes and Google API—that is, neural machine translation (NMT)—under the supervision of a professional translation trainer. As a result, accuracy is assured, and we may proceed to examine the naturalness of the subtitles.

The LB identification operations in this case study are carried out in two phases based on the structural and functional taxonomies (Biber, Conrad, and Cortes 2004, 379–382). In the first phase, three structural types of LBs, namely, bundles containing verb phrase fragments (e.g., *it's going to be*), bundles including dependent clause fragments (*I want you to*), and bundles incorporating noun and prepositional phrase fragments (*the end of the*) in the corpus, are manually annotated and calculated numerically, corresponding to the distribution of LBs across structural types in Biber, Conrad, and Cortes (2004, 379–383).

In the second phase, three primary functional classifications of LBs as listed in the classroom teaching from Biber, Conrad, and Cortes (2004) and academic lectures from Liu and Chen (2020b) are applied to examine the bundles in the MOOC MT-translated output: stance bundles, discourse organizing bundles, and referential bundles. Biber, Conrad, and Cortes compared the LBs in the main registers of classroom teaching and textbooks from the T2K-SWAL Corpus with those in conversation and academic prose and set a relatively high-frequency cut-off (40 times per million words to be included in the analysis, considering only

4 grams). Here, the corpus contains 176 texts on classroom teaching with 1,240,800 words in English.

Meanwhile, Liu and Chen (2020b) built a corpus to explore the frequent LBs in academic lectures following Biber, Conrad, and Cortes's taxonomy. The corpus on academic lectures consists of 565 lecture scripts from 71 courses delivered by 77 different teachers from various universities in the United States between 2004 and 2014, with data gathered from MIT OpenCourseWare, Open Yale Courses, and university channels on YouTube, all of which were recorded in regular university class sessions. There are 565 academic lectures in the AL corpus, which comprises approximately 4.38 million words evenly spread across the four disciplines. Both studies, by Biber, Conrad, and Cortes (2004) and Liu and Chen (2020b), were used to extract LBs from the corpus investigated in this case study. Simultaneously, we may be able to recommend which list can aid in the extraction of more LBs for use as a parameter not only for the naturalness of the MOOC subtitles but also as a reference for prospective adjustment.

Note that the corpora analyzed by Biber, Conrad, and Cortes (2004) and Liu and Chen (2020b) were not created specifically for online courses, such as the MOOCs reviewed in this case study. Their findings also support the argument for the problem of naturalness, which is expected to be explored in the MOOC subtitles and is still under-researched, particularly in audiovisual translation studies.

Moreover, in most of the previous corpus-based studies on LBs, 4 grams has always been justified as a valid and reliable length for multi-word sequences (e.g., Biber, Conrad, and Cortes 2004; Liu and Chen 2020a, 2020b; Wang 2017). In this case study, due to the shorter sentences, 3 to 4 grams in the MOOC MT-translated subtitles are used as criteria to extract as many LBs as feasible.

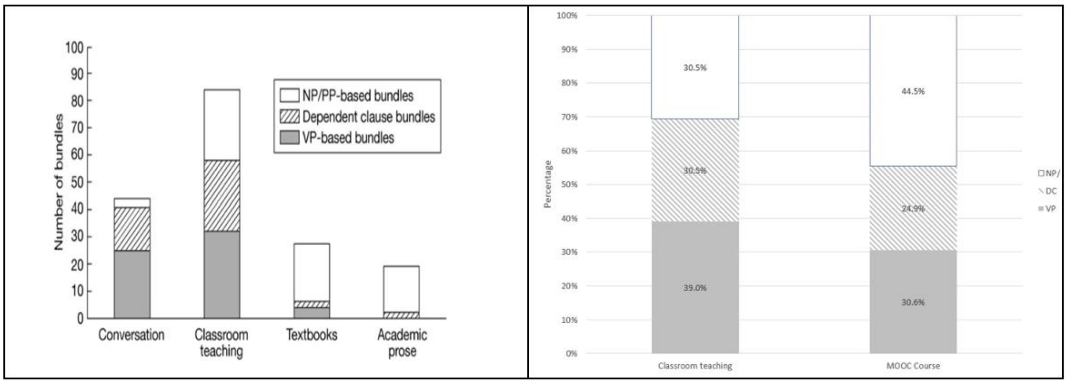
4. Results and Discussion

For the first phase of analysis, the distribution of the structural types across registers (VP-based, dependent clause, and NP/PP-based) was investigated throughout the translated subtitles in the MOOC. As shown in figure 1 (*left*), the corpus T2K-SWAL on which Biber, Conrad, and Cortes (2004) are based demonstrates that, in terms of the quantity of distinct LBs, classroom considerably outnumbers conversation, textbook, and academic prose. The number of NP/PP-based, dependent clause-based, and VP-based LBs in classroom teaching is rather

evenly distributed, with VP having a slightly greater number. Biber, Conrad, and Cortes concluded that classroom teaching depends on LBs linked with both spoken and written forms. Classroom teaching, like conversation, extensively includes VP-based LBs, noun phrase and prepositional phrase LBs, as in academic literature and textbooks. Thus, there is an unusually high density of LBs in classroom teaching, as this register heavily relies on both 'oral' and 'literate' bundles.

In contrast, from the MOOC translated corpus, it is discovered that NP-/PP-based clauses dominate the structural distribution across all LBs, whereas dependent clauses are the least numerous (figure 1, *right*). That is, the finding might be explained by the structural distribution of LBs in the MT output of MOOC subtitles tending toward the 'literate,' making it less 'oral.' According to Biber, Conrad, and Cortes (2004, 397), there is a substantial link between structural type and discourse function for LBs. Reduced dependent-sentence fragments may very likely lead to fewer stance bundles in the function distribution of LBs, which frequently represent the lecturer's attitude, obligation, or purpose. Similarly, a rise in noun phrases or prepositional phrase fragments would increase in referential bundles, which can relate to identification, specification, and time/place/text reference. These patterns are obviously closely connected to the register.

Figure 1. Distribution of LBs across structural types (*left*) and Biber, Conrad, and Cortes's Classroom Teaching and MOOC explored in this case study (*right*)



The three taxonomies of LBs in functional distribution in MOOC translated subtitles are extracted for the second phase of inquiry based on the lists supplied by Biber, Conrad, and Cortes (2004) and Liu and Chen (2020b). Table 1 shows the functional distribution of LBs in the small corpus, with the number in parenthesis indicating the occurrence. It is found that the

list provided by Biber, Conrad, and Cortes (2004) could extract more LBs in terms of number and type in the corpus explored in the present case study.

Based on the findings in phase 1 and the interrelationship between structural and functional distributions of LBs, it is expected that an increase in NP/PP-based phrases will result in an increase in referential bundles. According to Ying Wang (2017), a higher-than-expected number of reference bundles in lectures show a sharing of traits between lectures and more formal styles of academic writing. The current study also supports the idea that academic lectures, even if in the form of subtitles, have certain similarities with academic writing. Furthermore, fewer dependent sentence fragments, as observed in phrase one, are projected, resulting in fewer attitude bundles per the list provided by Liu and Chen (2020b). Among the stance bundles, it should be noted that the bundle *can be used* has been applied more than 20 times, still demonstrating the "impersonal" use (Biber, Conrad, and Cortes 2004, 386), which is common in academic writing. Given the near-complete lack of epistemic attitude comparing those in both lists, the academic register in this MOOC translated subtitles again leans towards the 'literate' side by obscuring the lecturer's epistemic standpoint.

Another noteworthy conclusion from table 1 is that discourse organizer bundles exceed the other two bundles by a large margin. In fact, the only functional category that employs all three structural forms is discourse organizers (Biber, Conrad, and Cortes 2004), and the translated subtitles of the MOOC appear to convey various discourse cues to help the audience "predict the nature of upcoming ideas and information" in academic speech (Nesi and Basturkmen 2006, 301). They also enable audience members to establish logical relationships between concepts (Kashiha and Chan 2014), which play similar roles in TED talks, as reported by Liu and Chen (2020b), to keep their audiences engaged and help them connect the dots between ideas more easily. Indeed, because the MOOCs platform is still in its early stages of development at this university, most of the courses recorded thus far comprise introductory courses from various academic areas. Therefore, the lecturers, similar to TED speakers, may try their best to use diverse discourse organizers, such as "overt signals" that a new topic is about to be introduced or a "preview" of the main topic (Biber and Barbieri 2007, 276), as demonstrated by two discourse organizers, such as *I want to* or *in other words*, are particularly common in usage (table 1).

Moreover, while Liu and Chen (2020b) believe that nonverbal approaches such as writing or drawing on blackboards or using props or tools may prevent academic lectures from using too many discourse organizers, the present case study argues that these nonverbal behaviors, to some extent, encourage the MOOC lecturer to apply more discourse organizers to demonstrate ideas and facilitate beginning learners' clear understanding.

Table 1. Functional distribution of LBs extracted from two lists

Lexical Bundles	Classroom Teaching **Biber, Conrad, and Cortes (2004)**	Academic Lecture Liu and Chen (2020**b**)
Stance Bundles Epistemic stance	more likely to (1)	x
Attitudinal/modality stance	if you want (1) you have to (1) it is important (1) it is necessary (4) I'm going to (6) be able to (2) can be used (22) it is possible (1)	you can see (5) you can see that (2) you want to (1) I'm going to (6)
Discourse Organizers Topic introduction/focus	to look at (1) take a look at (3) I want to (28) want to talk (5)	x
Topic elaboration/clarification	has to do with (1) as well as (3) on the other hand (9)	in other words (16) is the same as (4) the same as (5)
Referential Expressions Identification/focus	one of the (14)	this is the (13) is a very (1) is not a (5) this is a (5)
Specification of attributes	a lot of (11) the size of (4) in the form of (2) in the case of (3) as a result (5) in the absence of (1) in the presence of (1)	the sum of (1)
	at the same time (2) at the time (2) as shown in (2)	

Time/place/text reference	the end of (8) the beginning of (2) in the middle of (4)

The exploration of two phases in this case study reveals that the MT-translated MOOC subtitles lean towards the 'literate' rather than the 'oral' based on the distribution of LBs across structural and functional types. A relevant finding reported by Yu (2022) could be applied to provide a contrast. As one of the few studies on MOOC languages, Yu (2022) explored English-medium MOOC video lectures provided by Chinese universities, finding that the language tends to be informational, explicitly referential, and formally planned. She also pointed out that the closest register in her corpus could be academic prose, official documents, and press reportage.

However, in the present case study, the MT output, to some extent, keeps academic register attributes, such as the 'oral' bundles, with extra discourse organizers to improve beginning learners' understanding and convey cues throughout the lecture. As a result, this case study indicates that analyzing the naturalness of subtitles is crucial for quality control to evoke reality (Pérez-González 2007; Romero-Fresco 2009) through natural and spontaneous-sounding lectures for students, both locally and globally. A fundamental grasp of the register retained by the MT output is also necessary before any modifications are made. Exploring the distribution of LBs may thus serve as an excellent tool for generating a notion of the register under consideration.

On the other hand, compared to the research by Biber, Conrad, and Cortes (2004) and Liu and Chen (2020b), the LBs retrieved from the small corpus in this case study are limited, and the finding may not be able to stand for the distributions of LBs in other MOOC MT-translated subtitles. Nonetheless, this case study supports Łucja Biel's (2018) finding that MOOC MT-translated subtitles have fewer types of LBs in terms of functional distributions than non-translation. Indeed, in light of Biel's insight into the naturalness of formulaicity in translation, more MOOC translated English subtitles should be investigated, although the translation of MOOC subtitles from this current platform is still being worked on. A corpus could even be built in order to generalize the specific set of formulaic language (Danilaviciene, Horbacauskiene, and Kasperaviciene 2017) in translation to identify or be used as a metric of the naturalness of academic register.

Finally, some modifications to the existing too-literate register noticed in the MOOC translated subtitles in this case study are suggested. First, because the MT output was generated using a CAT tool, either of the lists—the one produced by Biber, Conrad, and Cortes (2004) or the one produced by Liu and Chen (2020b)—could be made into a parallel corpus with Chinese equivalence and used as translation memory (TM) prior to MT. Based on the conclusions of this case study, the list supplied by Biber, Conrad, and Cortes (2004) may extract more bundles from the corpus and, thus, should be prioritized. The idea of turning LBs into a parallel corpus and using them as TM is supported by the study conducted by Grabowski (2018), suggesting that the functionally-aligned bilingual LBs could be integrated into domain-specific glossaries or terminological datasets created by translators using CAT software. While it appears that only the SMT and RBMT systems were examined in the study (Grabowski 2018), this idea is thought to be applicable to the NMT system as well.

Of course, future researchers adopting this suggestion will face certain hurdles. As described by Grabowski (2018), it may not always be straightforward to discover an equivalent term in the source language, Chinese in this case study, for an LB since LBs may transmit changing pragmatic meanings depending on the context of use, posing a practical issue in translating. In his opinion, such challenges can only be rectified by giving extra meta-situational data because the accurate identification of meanings and functions of LBs ultimately affects the naturalness of translations, impacting total translation quality.

Secondly, it is suggested that the pre-editing of source text should be given more attention to preserve the academic register in the original. While studies on controlled language have emerged in the field of translation studies (e.g., Miyata 2021), there appears to be no substantial research, let alone training translators, on pre-editing the Chinese language. As a result, when it is necessary to pre-edit Chinese, translators may simply adhere to the criteria used in the source text, which may result in a significant change of controlled language before MT, as in this case, potentially leading to the significance of diminishing stance bundles. While Biel (2018) emphasized source language interference regarding the naturalness of formulaicity in translation, aligning all trainee translators throughout the pre-editing phase to the target language is highly recommended.

5. Conclusion

This case study is one of the few works investigating the naturalness of subtitles using MT. The investigation was divided into two phases based on the structural and functional attributes of LBs. It was discovered that the distribution of LBs across structural types in the corpus shifted the academic register toward a more 'literate' side, with NP/PP-based bundles outnumbering the other two. This inevitably entails a rise of referential bundles in the functional exploitation of LBs. Additionally, the high frequency of discourse organizers taken from two lists provided by Biber, Conrad, and Cortes (2004) and Liu and Chen (2020b) aids in retaining the 'oral' element of the academic register. Furthermore, the number of LBs extracted via the two lists was not as large as projected.

This discovery has two implications. First, it is suggested that more MOOC translated subtitles be explored in the future to build up a set of formulaic language per se for the parameter of naturalness; second, modifications of the academic register may be possible if Chinese equivalents for LBs are sought to integrate a parallel corpus for TM into the CAT tool for future use while also aligning all translators, particularly at the pre-editing stage. Overall, the present case study sheds some light on the naturalness of MOOC MT-translated subtitles via LBs.

References

Arenas, Ana Guerberof. 2019. "Pre-editing and Post-editing." In *The Bloomsbury Companion to Language Industry Studies*, edited by Erik Angelone, Maureen Ehrensberger-Dow, and Gary Massey, 333–360. London: Bloomsbury.

Arnon, Inbal, and Neal Snider. 2010. "More Than Words: Frequency Effects for Multi-word Phrases." *Journal of Memory and Language* 62 (1): 67–82. doi:10.1016/j.jml.2009.09.005.

Aston, Guy. 2018. "Acquiring the Language of Interpreters: A Corpus-based Approach." In *Making Way in Corpus-based Interpreting Studies*, edited by Mariachiara Russo, Claudio Bendazzoli, and Bart Defrancq, 83–96. Singapore: Springer.

Atapattu, Thushari, and Katrina Falkner. 2018. "Impact of Lecturer's Discourse for Students' Video Engagement: Video Learning Analytics Case Study of MOOCs." *Journal of Learning Analytics* 5 (3): 182–197. doi:10.18608/jla.2018.53.12.

Berūkštienė, Donta. 2017. "A Corpus-driven Analysis of Structural Types of Lexical Bundles in Court Judgments in English and Their Translation into Lithuanian." *Kalbotyra* 70:7–31. doi:10.15388/Klbt.2017.11181.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

———. 2006a. "Stance in Spoken and Written University Registers." *Journal of English for Academic Purposes* 5 (2): 97–116. doi:10.1016/j.jeap.2006.05.001.

———. 2006b. *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

Biber, Douglas, and Federica Barbieri. 2007. "Lexical Bundles in University Spoken and Written Registers." *English for Specific Purposes* 26 (3): 263–286. doi:10.1016/j.esp.2006.08.003.

Biber, Douglas, and Susan Conrad. 2019. *Register, Genre, and Style*. 2nd ed. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad, and Geoffrey Leech. 2002. *Longman Student Grammar of Spoken and Written English*. Harlow: Pearson.

Biber, Douglas, Susan Conrad, and Viviana Cortes. 2004. "*If You Look at...*: Lexical Bundles in University Teaching and Textbooks." *Applied Linguistics* 25 (3): 371–405. doi:10.1093/applin/25.3.371.

Biel, Łucja. 2018. "Lexical Bundles in EU Law: The Impact of Translation Process on the Patterning of Legal Language." In *Phraseology in Legal and Institutional Settings: A Corpus-based Interdisciplinary Perspective*, edited by Stanisiaw Goźdź-Roszkowski and Gianluca Pontrandolfo, 11–26. London: Routledge.

Bogucku, Łukasz. 2019. *Areas and Methods of Audiovisual Translation Research*. Bern: Peter Lang.

Chaume-Varela, Frederic. 2007. "Quality Standards in Dubbing: A Proposal." *TradTerm* 13:71–89. doi:10.11606/issn.2317-9511.tradterm.2007.47466.

Chiaro, Delia. 2008. "Issues of Quality in Screen Translation: Problems and Solutions." In *Between Text and Image: Updating Research in Screen Translation*, edited by Delia Chiaro, Christine Heiss, and Chiara Bucaria, 241–256. Amsterdam: John Benjamins.

Cortes, Viviana. 2004. "Lexical Bundles in Published and Student Disciplinary Writing: Examples from History and Biology." *English for Specific Purposes* 23 (4): 397–423. doi:10.1016/j.esp.2003.12.001.

Coxhead, Averil, Dang, Thi Ngoc Yen, and Shota Mukai. 2017. "Single and Multi-word Unit Vocabulary in University Tutorials and Laboratories: Evidence from Corpora and Textbooks." *Journal of English for Academic Purposes* 30:66–78. doi:10.1016/j.jeap.2017.11.001.

Crawford Camiciottoli, Belinda. 2007. *The Language of Business Studies Lectures: A Corpus-assisted Analysis*. Amsterdam: John Benjamins.

Csomay, Eniko. 2006. "Academic Talk in American University Classroom: Crossing the Boundaries of Oral-literate Discourse?" *Journal of English for Academic Purposes* 5 (2): 117–135. doi:10.1016/j.jeap.2006.02.001.

———. 2007. "A Corpus-based Look at Linguistic Variation in Classroom Interaction: Teacher Talk Versus Student Talk in American University Classes." *Journal of English for Academic Purposes* 6 (4): 336–355. doi:10.1016/j.jeap.2007.09.004.

Danilaviciene, Greta, Jolita Horbacauskiene, and Ramune Kasperaviciene. 2017. "On Formulaic Language in Subtitling and Voice-over." *Translationes* 9 (1): 71–82. doi:10.1515/tran-2017-0004.

Das, Ananda, and Partha Pratim Das. 2019. "Automatic Semantic Segmentation and Annotation of MOOC Lecture Videos." In *Digital Libraries at the Crossroads of Digital Information for the Future*, edited by Adam Jatowt, Akira Maeda, and Sue Yeon Syn, 181–188. Berlin: Springer.

Días-Cintas, Jorge. 2019. "Audiovisual Translation." In *The Bloomsbury Companion to Language Industry Studies*, edited by Erik Angelone, Maureen Ehrensberger-Dow, and Gary Massey, 209–230. London: Bloomsbury.

Doherty, Stephen. 2017. "Issues in Human and Automatic Translation Quality Assessment." In *Human Issues in Translation Technology*, edited by Dorothy Kenny, 131–148. London: Routledge.

Doherty, Stephen, and Jan-Louis Kruger. 2018. "The Development Eye Tracking in Empirical Research on Subtitling and Captioning." In *Seeing into Screens. Eye Tracking and the Moving Image*, edited by Tessa Dwyer, Claire Perkins, Sean Redmond, and Jodi Sita, 46–64. London: Bloomsbury.

Ellis, Nick C., and Rita Simpson-Vlach. 2009. "Formulaic Language in Native Speakers: Triangulating Psycholinguistics, Corpus Linguistics, and Education." *Corpus Linguistics and Linguistic Theory* 5 (1): 61–78. doi:10.1515/CLLT.2009.003.

Ellis, Nick C., Rita Simpson-Vlach, and Carson Maynard. 2008. "Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL." *TESOL Quarterly* 42 (3): 375–396. doi:10.1002/j.1545-7249.2008.tb00137.x.

Georgakopoulou, Panayota (Yota). 2019. "Technologization of Audiovisual Translation." In *The Routledge Handbook of Audiovisual Translation*, edited by Luis Pérez-González, 516–539. London: Routledge.

Grabowski, Łukasz. 2018. "On Identification of Bilingual Lexical Bundles for Translation Purposes: The Case of an English-Polish Comparable Corpus of Patient Information Leaflets." In *Multiword Units in Machine Translation and Translation Technology*, edited by Ruslan Mitkov, Joanna Monti, Gloria Corpas Pastor, and Violeta Seretan, 181–200. Amsterdam: John Benjamins.

Hyland, Ken. 2008. "Academic Clusters: Text Patterning in Published and Postgraduate Writing." *International Journal of Applied Linguistics* 18 (1): 41–62. doi:10.1111/j.1473-4192.2008.00178.x.

Kashiha, Hadi, and Swee Heng, Chan. 2014. "Discourse Functions of Formulaic Sequences in Academic Speech Across Two Disciplines." *GEMA: Online Journal of Language Studies* 14 (2): 15–27. doi:10.17576/GEMA-2014-1402-02.

Kuo, Arista Szu-Yu. 2020. "The Tangled Strings of Parameters and Assessments in Subtitling Quality: An Overview." In *The Palgrave Handbook of Audiovisual Translation and Media Accessibility*, edited by Łukasz Bogucki and Mikołaj Deckert, 437–458. Cham: Palgrave Macmillan.

Lee, Changsoo. 2013. "Using Lexical Bundle Analysis as Discovery Tool for Corpus-based Translation Research." *Perspectives: Studies in Translatology* 21 (3): 378–395. doi:10.1080/0907676X.2012.657655.

Lee, Joseph J., and Nicholas C. Subtirelu. 2015. "Metadiscourse in the Classroom: A Comparable Analysis of EAP Lessons and University Lectures." *English for Specific Purposes* 37:52–62. doi:10.1016/j.esp.2014.06.005.

Li, Jie, and Nobert Schmitt. 2009. "The Acquisition of Lexical Phrases in Academic Writing: A Longitudinal Chunks." *Modern Foreign Languages* 39 (2): 246–256. doi:10.1016/j.jslw.2009.02.001.

Liu, Chen-Yu, and Hao-Jan Howard Chen. 2020a. "Analyzing the Function of Lexical Bundles in Undergraduate Academic Lectures for Pedagogical Use." English for Specific Purposes 58:122–137. doi:10.1016/j.esp.2019.12.003.

———. 2020b. "Functional Variation of Lexical Bundles in Academic Lectures and TED Talks." *Register Studies* 2 (2): 176–208. doi:10.1075/rs.18003.liu.

Liu, Kanglong, Joyce Oiwun Cheung, and Riccardo Moratto. 2022. "Lexical Bundles in Fictional Dialogues of Two *Hongloumeng* Translations: A Corpus-assisted Approach." In *Advances in Corpus Applications in Literary and Translation Studies*, edited by Riccado Moratto and Defeng Li, 229–253. London: Routledge.

Liu, Kanglong, and Muhammad Afzaal. 2021. "Translator's Style Through Lexical Bundles: A Corpus-driven Analysis of Two English Translations of *Hongloumeng*." *Frontiers in Psychology* 12. doi:10.3389/fpsyg.2021.633422.

Li, Yang, and Sandra L. Halverson. 2020. "A Corpus-based Exploration into Lexical Bundles in Interpreting." *Across Languages and Cultures* 21 (1): 1–22. doi:10.1556/084.2020.00001.

———. 2022. "Lexical Bundles in Formulaic Interpreting: A Corpus-based Descriptive Exploration." *Translation and Interpreting Studies* 19 (2): 33–56. doi:10.1075/tis.19037.li.

Miyata, Rei. 2021. *Controlled Document Authoring in a Machine Translation Age*. London: Routledge.

Neely, Elizabeth, and Viviana Cortes. 2009. "A Little Bit About: Analyzing and Teaching Lexical Bundles in Academic Lectures." *Language Value* 1 (1): 17–38. http://www.e-revistes.uji.es/languagevalue.

Nesi, Hilary, and Helen Basturkmen. 2006. "Lexical Bundles and Discourse Signalling in Academic Lectures." *International Journal of Corpus Linguistics* 11 (3): 283–304. doi:10.1075/ijcl.11.3.04nes.

Nikolic, Kristijan. 2015. "The Pros and Cons of Using Templates in Subtitling." In *Audiovisual Translation in a Global Context. Mapping and Ever-changing Landscape*, edited by Rocío Baños Piñero and Jorge Díaz Cintas, 192–202. Basingstoke: Palgrave Macmillan.

O'Brien, Sharon. 2017. "Machine Translation and Cognition." In *The Handbook of Translation and Cognition*, edited by John W. Schwieter and Aline Ferreira, 313–331. New Jersey: Wiley Blackwell.

Pavesi, Maria. 2008. "Spoken Language in Film Dubbing: Target Language Norms, Interference and Translational Routines." In *Between Text and Image. Updating Research in Screen Translation*, edited by Delia Chiaro, Christine Heiss, and Chiara Bucaria, 79–99. Amsterdam: John Benjamins.

———. 2009. "Dubbing English into Italian: A Closer Look at the Translation of Spoken Language." In *New Trends in Audiovisual Translation*, edited by Jorge Díaz-Cintas, 197–209. Toronto: Multilingual Matters.

———. 2013. "*This* and *That* in the Language of Film Dubbing: A Corpus-based Analysis." *Meta* 58 (1): 103–133. doi:10.7202/1023812ar.

Pavesi, Maria, Maicol Formentelli, and Elisa Ghia. 2014. *The Languages of Dubbing. Mainstream Audiovisual Translation in Italy*. Bern: Peter Lang.

Pedersen, Jan. 2017. "The FAR Model: Assessing Quality in Interlingual Subtitling." *The Journal of Specialised Translation*, no. 28, 210–219. https://jostrans.soap2.ch/issue28/art_pedersen.php.

Perego, Elisa. 2016. "History, Development, Challenges and Opportunities of Empirical Research in Audiovisual Translation." *Across Languages and Cultures* 17 (2): 155–162. doi:10.1556/084.2016.17.2.1.

Pérez-González, Luis. 2007. "Appraising Dubbed Conversation. Systemic Functional Insights into the Construal of Naturalness in Translated Film Dialogue." *The Translator* 13 (1): 1–38. doi:10.1080/13556509.2007.10799227.

Pettit, Zoë. 2005. "Translating Register, Style and Tone in Dubbing and Subtitling." *The Journal of Specialised Translation*, no. 4: 49–65. https://jostrans.soap2.ch/issue04/art_pettit.php.

Plevoets, Koen, and Bart Defrancq. 2018. "The Cognitive Load of Interpreters in the European Parliament: A Corpus-based Study of Predictors for the Disfluency *uh(m)*." *Interpreting* 20 (1): 1–32. doi:10.1075/intp.00001.ple.

Romero-Fresco, Pablo. 2009. "Naturalness in the Spanish Dubbing Language: A Case of Not-so-close Friends." *Meta* 54 (1): 49–72. doi:10.7202/029793ar.

———. 2011. *Subtitling through Speech Recognition: Respeaking*. Manchester: St. Jerome.

Salazar, Danica. 2011. "Lexical Bundles in Scientific English: A Corpus-based Study of Native and Non-native Writing." PhD diss., University of Barcelona.

Sharma, Kshitij, Sarah D'Angelo, Darren Gergle, and Pierre Dillenbourg. 2016. "Visual Augmentation of Deictic Gestures in MOOC Videos." Paper presented at 12th International Conference of the Learning Sciences, Singapore, June 20–24.

Shi, Jianwei, Christian Otto, Anett Hoppe, Peter Holtz, and Ralph Ewerth. 2019. "Investigating Correlations of Automatically Extracted Multimodal Features and Lecture Video Quality." In *Proceedings of the 1st International Workshop on Search as Learning with Multimedia Information*, 11–19. https://dl.acm.org/doi/10.1145/3347451.3356731.

Simpson, Rita. 2004. "Stylistic Features of Academic Speech: The Role of Formulaic Expressions." In *Discourse in the Professions: Perspectives from Corpus Linguistics*, edited by Ulla Connor and Thomas A. Upton, 37–64. Amsterdam: John Benjamins.

Sinclair, John, ed. 1995. *Collins Cobuild English Dictionary*. Glasgow: HarperCollins.

Tremblay, Antoine, Bruce Derwing, Gary Libben, and Chris Westbury. 2011. "Processing Advantages of Lexical Bundles: Evidence from Self-paced Reading and Sentence Recall Tasks." *Language Learning* 61 (2): 569–613. doi:10.1111/j.1467-9922.2010.00622.x.

Uchidiuno, Judith Odili, Amy Ogan, Evelyn Yarzebinski, and Jessica Hammer. 2018. "Going Global: Understanding English Language Learner's Student Motivation in English-language MOOCs." *International Journal of Artificial Intelligence in Education* 28 (4): 528–552. doi:10.1007/s40593-017-0159-7.

Valdéon, Roberto A. 2008. "Inserts in Modern Script-writing and Their Translation into Spanish." In *Between Text and Image. Updating Research in Screen Translation*, edited by Delia Chiaro, Christine Heiss, and Chiara Bucaria, 117–132. Amsterdam: John Benjamins.

Vorobyeva, Alexandra. 2018. "Language Acquisition Through Massive Open Online Courses (MOOCs), Opportunities and Restrictions in Educational University Environment." *XLinguae* 11 (2): 136–146. doi:10.18355/XL.2018.11.02.11.

Wang, Pei-Yu, Miao-Chin Chiu, and Yu-Tzu Lee. 2020. "Effects of Video Lecture Presentation Style and Questioning Strategy on Learning Flow Experience." *Innovations in Education and Teaching International* 58 (4): 473–483. doi:10.1080/14703297.2020.1754272.

Wang, Wei, Lihuan Guio, Ling He, and Yenchun Jim Wu. 2019. "Effects of Social-Interactive Engagement on the Dropout Ratio in Online Learning: Insights from MOOC." *Behaviour & Information Technology* 38 (6): 621–636. doi:10.1080/0144929X.2018.1549595.

Wang, Ying. 2017. "Lexical Bundles in Spoken Academic ELF." *International Journal of Corpus Linguistics* 22 (2): 187–211. doi:10.1075/ijcl.22.2.02wan.

Wood, David. 2010. "Lexical Clusters in an EAP Textbook Corpus." In *Perspectives on Formulaic Language: Acquisition and Communication*, edited by David Wood, 88–106. London: Continuum.

Wray, Allison, and Mick Perkins. 2000. "The Functions of Formulaic Language: An Integrated Model." *Language and Communication* 20 (1): 1–28. doi:10.1016/S0271-5309(99)00015-4.

Yu, Xiaoli. 2022. "A Multi-dimensional Analysis of English-medium Massive Open Online Courses (MOOCs) Video Lectures in China." *Journal of English for Academic Purposes* 55:1–14. doi:10.1016/j.jeap.2021.101079.