Journal of Audiovisual Translation

Human vs. Machine-Assisted Subtitle Translation in MOOCs: A Corpus-Based Case Study

Tzu-yi Elaine Lee ✉
Chung Yuan Christian University

Citation: Lee, Tzu-yi. (2025). Human vs. Machine-Assisted Subtitle Translation in MOOCs: A Corpus-Based Case Study. Journal of Audiovisual Translation, 8(2), 1–16.
https://doi.org/10.47476/jat.v8i2.2025.357

Editor(s): D. Chiaro & L. Rossato

Received: November 6, 2024

Published: December 9, 2025

Abstract

Massive Open Online Courses (MOOCs) have become important audiovisual resources in higher education, offering accessible, high-quality learning materials to a global audience. While MOOCs have been extensively studied, relatively little attention has been paid to the language used in MOOC lectures, particularly from the perspectives of audiovisual translation (AVT) and machine translation (MT). This study investigates lexical bundles in machine-translated and human-translated corpora from four MOOCs through a case study approach, using AntConc 4.3.1 to extract frequently recurring bundles categorized by structure and function. Findings suggest that MT-generated translations tend to align more closely with a "literate" register, dominated by referential bundles, whereas human-translated subtitles reflect a more "oral" register, with discourse organizers comprising nearly 40% of the total. While these results diverge from Biber and Barbieri's (2007) conclusions, they are consistent with studies indicating that MT output resembles academic lectures, while human translations show features similar to those found in general TED Talk discourse. Moreover, despite being produced at different times, both corpora include exact matches and comparable lexical bundles across various instructional stages. This study offers preliminary insights into the academic register of MOOC subtitle translation and contributes to the growing body of research in audiovisual translation.

Key words: MOOCs, lexical bundles, MT-assisted translation, human translation, audiovisual translation, literate and oral register.

✉ bluestone1207@gmail.com, http://orcid.org/0000-0002-1139-9294

Introduction

The increasing reliance on audiovisual translation (AVT) in educational contexts has brought new attention to the linguistic characteristics of Massive Open Online Courses (MOOCs), particularly the subtitles that facilitate their global reach. MOOCs have transformed higher education by offering multimodal, freely accessible learning materials that integrate spoken, visual, and textual input (Uchidiuno et al., 2018; Wang et al., 2020). As audiovisual products, MOOCs necessitate translation strategies that align with both their pedagogical objectives and their multimodal constraints, particularly when subtitles are generated through machine translation (MT) or created by human translators. Despite their educational relevance, the language used in MOOC subtitles—especially in terms of how different translation modes shape discourse—is still underexamined (Yu, 2022).

This study investigates lexical bundles (LBs) in human-translated and MT-generated subtitles from four MOOC courses to better understand how translation mode influences register variation in subtitled educational discourse. Rather than focusing on changes over time, the analysis examines stylistic contrasts between translation types, providing a case study on how translation workflows influence the textual realization of pedagogical content. In particular, the study highlights how subtitling as a form of intralingual and intersemiotic mediation (cf. Díaz Cintas & Remael, 2021; Pavesi, 2019) can affect the interpersonal and informational tone of instructional discourse, with implications for translator decision-making and learner engagement.

Lexical bundles—defined by Biber et al. (2002, p. 190) as "recurrent expressions"—are essential indicators of register, especially in spoken academic genres. Although they lack the idiomaticity of fixed phrases or the structural complexity of full clauses, LBs function as discourse organizers, stance markers, and referential scaffolds in classroom teaching and academic writing (Biber et al., 2004; Hyland, 2008). Their high frequency in educational settings makes them particularly useful for analyzing how translated subtitles preserve or modify pedagogical tone, especially in multimodal contexts where space, timing, and information density are tightly constrained. Previous research has explored the role of LBs in academic lectures, TED Talks, and textbooks (Biber et al., 2004; Liu & Chen, 2020a, 2020b; Wang, 2017). These studies suggest that spoken academic registers tend to favor bundles that signal topic shifts, speaker stance, and referential clarity. Classroom discourse, for instance, blends features of both spoken and written language, with heavy use of noun/prepositional phrase bundles (NP/PP) alongside verb-based stance expressions. Liu and Chen (2020b) compared LBs in TED Talks and lectures, showing how discourse structure and audience orientation influence bundle selection. However, few studies have examined how these patterns are reshaped through translation, particularly when the subtitles are generated through automated or semi-automated translation workflows.

MOOCs, by design, combine elements of classroom teaching, public communication, and interactive media. Their subtitles—often derived from condensed transcripts and then translated—are an ideal site for examining how human and machine translation workflows yield different linguistic realizations of the same instructional content. MT-assisted translations often involve pre-editing and

post-editing, which can increase consistency but also introduce formulaic or written-like phrasing (Grabowski, 2018; Lee, 2024). By contrast, human translators may prioritize intonation, engagement, or oral fluency, reflecting different assumptions about audience needs and register expectations.

This study focuses on four MOOC courses—two translated by humans and two translated using MT workflows from Mandarin to English—and addresses the following research questions:

1. What are the most frequent lexical bundles in human-translated and MT-generated English subtitles of MOOCs, particularly in terms of their distribution across structural and functional taxonomies?
2. How do the lexical bundles in both sets compare, and what are the implications for future audiovisual translation of MOOC subtitles?

In this context, corpus linguistics offers powerful tools for identifying recurring phraseological patterns and uncovering the register-transforming effects of translation mode. As Pavesi (2019), Bruti (2025), and Pérez-González (2014) argue, corpus-based methodologies are essential in AVT research for tracing variation in discourse structure, interpersonal tone, and functional phrasing. These methods support both empirical pattern detection and interpretive discourse analysis. The present study adopts this approach by employing lexical bundle analysis to examine how various translation workflows influence register variation in MOOC subtitle discourse. By combining frequency-based extraction with functional categorization, the analysis bridges the gap between corpus description and AVT application, contributing to current research on register realization in subtitled educational media (cf. Biber et al., 2004; Ädel & Erman, 2012; Bruti, 2020).

By situating MOOC subtitles as multimodal AVT products and applying corpus-based methods to their analysis, this study contributes to a deeper understanding of how translation mode shapes register and pedagogical tone in subtitled educational discourse.

1. Corpus and Methodology

This case study analyzes four MOOC courses: *Introduction to Medical Devices and Principles*, *Decryption of Biomaterials*, *Economics Everywhere*, and *Statistics: Let the Numbers Speak*. All courses were originally delivered in Chinese and are hosted by a university in northern Taiwan. The first two courses are taught by faculty members from the Department of Biomedical Engineering, while the latter two are taught by instructors from the Departments of International Business and Information Management, respectively. Each course spans six to eight hours and is divided into 20 to 50 video segments, each approximately 15 minutes in length. Each video produced a corresponding set of Chinese subtitles, which served as the source for English translation. Accordingly, the number of videos corresponds to the number of texts analyzed. Two of the courses were translated into English by professional human translators. The other two involved a hybrid workflow that incorporated human translators, the computer-assisted translation (CAT) tool *Termsoup*, and Google's neural machine translation (NMT) API. The MOOC platform at this Taiwanese university is still in its early

stages. Faculty members have been encouraged to produce video-based courses for both local and international audiences, particularly in response to the shift toward online education during the COVID-19 pandemic. To broaden accessibility, the lectures were transcribed and translated into English subtitles. Translators adopted varied approaches in the subtitling process: some followed a manual, line-by-line strategy, while others employed CAT or MT tools to support consistency and efficiency. These workflows reflect a range of translation practices rather than adherence to a standardized procedure. The translators were commissioned by Cloud Academy, the unit responsible for the production and dissemination of the university's MOOC content.

For MT-assisted translations, a pre-editing process was first applied to the Chinese transcripts to standardize terminology, simplify syntax, and enhance consistency of the source text. These transcripts were then translated using Google's MT API and post-edited via the *Termsoup* CAT interface. Although the initial output was generated by MT, it was subsequently refined through CAT-based post-editing. In this study, this hybrid workflow is referred to as CAT/MT-assisted translation to acknowledge its blended nature. Human-translated subtitles, by contrast, were produced independently by professional translators without the aid of MT tools. All translations—regardless of method—underwent professional review to ensure linguistic accuracy and overall coherence. Table 1 summarizes the basic information of the four MOOC courses included in this study.

Table 1

MOOCs Course Information

Course Title	Number of Videos (Duration)	Translation Method	Translated Word Count	Translation Year
Introduction to Medical Devices and Principles	32 (6h 32m 59s)	MT-Generated	81,536	2021-2022
Decryption of Biomaterials	25 (6h 28m 59s)	MT-Generated	39,773	2021-2022
Economics Everywhere	46 (8h 54m 27s)	Human Translation	81,720	2020-2021
Statistics: Let the Numbers Speak	52 (6h 8m 13s)	Human Translation	48,567	2020-2021

Source. Author, based on data from Cloud Academy, accessed Jan 2025.

The identification of LBs in this study was carried out in two distinct phases, following the structural and functional taxonomies proposed by Biber et al. (2004, pp. 379–382). In the first phase, three

structural categories were manually annotated and quantified across the corpus: verb phrase (VP) fragments (e.g., “it’s going to be”), dependent clause (DC) fragments (e.g., “I want you to”), and noun or prepositional phrase (NP/PP) fragments (e.g., “the end of the”). The second phase involved assigning functional classifications to the extracted bundles, based on a framework originally developed by Biber et al. (2004) for analyzing classroom discourse. These functions include stance bundles, discourse-organizing bundles, and referential bundles, and were applied to assess the functional roles of bundles across both corpora, with particular attention to the MT-generated subtitles. It is worth noting that structural and functional classifications do not align in a one-to-one manner. While certain structural types frequently correspond to specific functions—for instance, NP/PP fragments often serve referential purposes—many bundles are multifunctional and context-dependent. For example, a VP-based bundle, such as “the one I want to show you”, may function either as a stance marker or as a discourse organizer, depending on its placement within the instructional sequence.

In their comparison of LBs across various registers—including classroom teaching, textbooks, conversation, and academic prose—Biber et al. (2004) relied on data from the T2K-SWAL Corpus (TOEFL 2000 Spoken and Written Academic Language Corpus), which was designed to represent the spoken and written registers encountered by university students in the US. Specifically, their classroom teaching data were collected from four academic institutions across six major disciplines. They applied a high-frequency cut-off of 40 occurrences per million words and focused exclusively on 4-gram sequences. Their corpus included 176 texts totaling 1,248,800 words drawn from live classroom instruction. In contrast, the present study analyzes translated subtitles from MOOC videos—a distinct modality with different production processes and audience profiles. Nevertheless, the consistent use of 4-gram sequences—combinations of four words that appear together more frequently than would be expected by chance—in corpus-based research has demonstrated its utility for identifying multi-word expressions (e.g., Biber et al., 2004; Liu & Chen, 2020a, 2020b; Wang, 2017). Accordingly, this study also adopts the 4-gram standard to ensure comparability with previous studies, while focusing on the specific context of subtitled MOOCs. LBs were extracted using AntConc 4.3.1 (Anthony, 2024), with the minimum n-gram size set at four words. Both frequency cut-off values and distributional criteria are acknowledged to be somewhat arbitrary (Biber & Barbieri, 2007). In this study, the frequency threshold was set to a minimum of two occurrences, following a similar approach in small-scale corpus analyses (e.g., Wu, 2021). Given the relatively limited size of the present corpus and the compressed nature of subtitle text, this threshold allows for the retrieval of salient yet less frequent lexical bundles. To reduce the influence of translator-specific usage patterns, each LB was also required to occur in at least five different subtitle files. This dual criterion—minimum frequency and cross-textual distribution—aims to balance representativeness with inclusivity, aligning with strategies commonly used in compact corpora (Biber et al., 2004; Cortes, 2004; Grabowski, 2018; Jalali et al., 2015).

2. Results and Discussion

This section addresses the two research questions posed in the study and presents the findings in relation to them. The analysis focuses on the most frequently occurring LBs identified in the two corpora—human-translated and machine-translated English subtitles of MOOCs—using the structural and functional framework proposed by Biber et al. (2004). Particular attention is given to the distinctive distributional patterns in each corpus. Where relevant, the findings are also compared with those reported for the classroom teaching register in Biber et al. (2004). Although LBs have been extensively studied in academic writing (e.g., Cortes, 2004; Hyland, 2008; Salazar, 2011), fewer studies have conducted in-depth analyses of spoken lecture discourse, particularly in relation to specific disciplinary contexts (Crawford Camiciottoli & Querol-Julián, 2016, p. 318). While some research has focused on academic spoken discourse (e.g., Crawford Camiciottoli, 2007), such work typically lacks comprehensive inventories of LBs that would allow for comparison across disciplines, such as those explored in the present study. The second part of this section compares LBs across the two corpora, with a focus on exact matches and functionally similar expressions. Finally, the implications of these findings for the translation of MOOC subtitles are discussed, following a summary of the key results below.

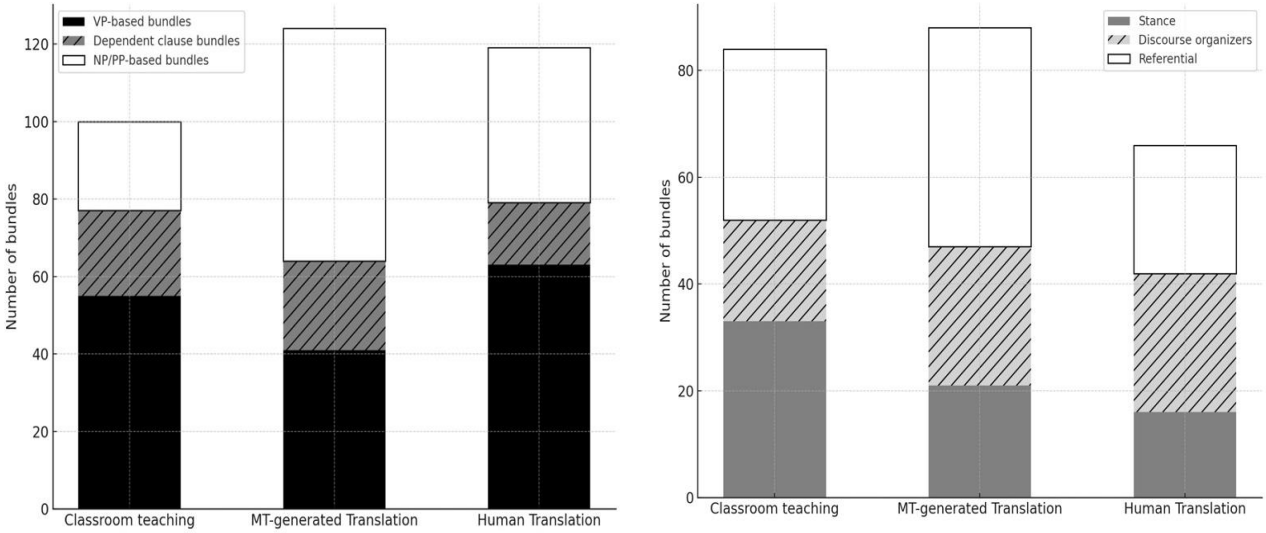
The number of LBs identified in the MT-generated translation corpus (100 bundles) is smaller than that found in the human-translated corpus (150 bundles). The most frequently occurring LB in the MT corpus is “in the human body”, which appears 44 times across 23 videos. This recurrence likely reflects the frequent reference to anatomical content in the course materials. The second most frequent bundle is “see you next time”, occurring 28 times in 27 videos. Its frequency suggests a conventional use as a closing formula in online lecture formats, rather than a marker of register-specific discourse. Three other bundles share the third-highest frequency, each appearing in 25 videos: “introduction to medical devices, medical devices and principles”, and “introduction to medical devices and principles”. All of these bundles belong to the category of noun/prepositional phrase fragments, which are commonly used in academic discourse to convey referential meaning by linking ideas to entities, locations, or objects. Their recurrence likely reflects the course titles, both of which introduce biomedical devices and biomaterials. It is also possible that the prevalence of these bundles was influenced by the pre-editing process, which may have standardized phrasing in the highly condensed original Chinese transcripts.

In the human-translated corpus, the LB *“*when it comes to” occurs most frequently, appearing 46 times across 28 videos. This bundle is commonly used in both spoken and written discourse to introduce a dependent clause, signaling a topic shift or the introduction of a new aspect for discussion. It thus serves a referential and explanatory function by directing attention towards a specific topic. Its frequent use may suggest that the translators were sensitive to moments in the lecture where topic transitions occurred. Compared to the MT-generated corpus—where LBs were more evenly distributed across videos, aside from the recurrent *“*in the human body”—the human-translated corpus contains several bundles that appear with high frequency but are concentrated in a limited number of videos. For example, *“*is equal to the” appears 35 times across only 7 videos;

"in the short run" appears 23 times in 7 videos; and "is equal to zero" occurs 20 times in 6 videos. These patterns may reflect the formulaic and repetitive nature of subject-specific instructional content, particularly in courses on statistics and economics. In such contexts, the recurrence of domain-specific bundles likely stems from content repetition rather than from translator-specific stylistic choices. However, it is also possible that translators rendered these recurring phrases literally without modification or rephrasing, particularly in workflows assisted by MT. The following section presents the structural and functional attributes of each corpus.

Figure 1

Structural and Functional Distribution of Lexical Bundles in MOOC Subtitles



Source. Author, based on classification framework from Biber et al. (2004).

Figure 1 presents a comparison of LBs extracted from MT-generated translations, human translations, and classroom teaching (Biber et al., 2004), with a focus on their structural and functional attributes as illustrated in two figures. Although different analytical tools were used, the relatively small corpora in this study contain LBs comparable to those found in Biber et al.'s large-scale classroom corpus, suggesting a dense distribution of bundles in translated subtitles. Beginning with their structural characteristics, LBs are categorized according to verb phrase fragments, dependent clause fragments, and noun or prepositional phrase components. According to Biber et al. (2004), the classroom teaching register draws on bundles common to both spoken and written registers. It frequently includes declarative and interrogative clause fragments (i.e., dependent clause bundles), as well as noun and prepositional phrase bundles similar to those found in academic prose and textbooks. Thus, the classroom teaching register reflects features of both "oral" and "literate" discourse. Turning to the current study, Figure 1 shows that AntConc 4.3.1 successfully extracted LBs that are frequent, structurally analyzable, and distributed across the subtitles, despite the relatively

small size of each corpus. The distribution of structural types differs slightly between the MT-generated and human-translated corpora—particularly in the relative presence of verb phrases, dependent clauses, and noun/prepositional phrases. Notably, noun and prepositional phrase bundles constitute a large portion of the total in the MT-generated translations, whereas verb clause fragments are more prevalent in the human-translated corpus. This contrast suggests that MT-generated translations tend to align more closely with a "literate" register, while human translations reflect more characteristics of spoken discourse. This tendency is consistent with Lee's (2024) findings, which likewise observed that MT-generated subtitles resembled written academic language. Two factors may account for this pattern. First, MT engines are typically trained on written corpora such as news articles, technical documentation, and manuals. Second, the pre-editing and post-editing stages involved in the MT workflow often condense the original content, removing repetition and rephrasing. By contrast, human-translated subtitles were produced without such modifications, which may explain the retention of oral features such as repetition and topic-initiation phrases.

Building on the structural classification discussed above, LBs can also be categorized according to their discourse functions, typically falling into three overarching types. Stance bundles express attitudes, evaluations, or degrees of certainty in relation to a given proposition. Discourse-organizing bundles contribute to textual coherence by signaling relationships between preceding and upcoming segments of discourse. Referential bundles, by contrast, make explicit reference to physical or abstract entities or to elements within the text itself, either by identifying the referent or by emphasizing a salient characteristic. Each of these functional categories encompasses multiple subcategories, each associated with more specialized discourse roles.

According to the functional taxonomy proposed by Biber et al. (2004), the right-hand figure in Figure 1 shows that the MT-generated corpus contains more lexical bundles than the human-translated corpus. Notably, it also includes a greater number of functional bundles than those identified in classroom teaching, based on Biber et al.'s (2004) larger corpus. The differences become more pronounced when functional categories are compared. While stance and referential bundles are balanced in the classroom teaching register, the MT-generated corpus exhibits a clear predominance of referential bundles. Specifically, the number of referential bundles in the MT corpus is nearly twice that of stance bundles. This pattern corresponds with the "literate" register often associated with MT-generated translations, which tend to contain a high proportion of NP/PP-based bundles serving referential functions (Biber et al., 2004). Although structural and functional categories do not map one-to-one, prior research shows a strong tendency for stance bundles to co-occur with verb phrase fragments, and for referential bundles to appear in the form of noun or prepositional phrases. These patterns in the classroom teaching corpus reflect its hybrid nature—combining spoken features such as evaluative stance with written, information-heavy functions typical of academic discourse. Interestingly, the present findings diverge from those of Biber and Barbieri (2007), who observed that stance bundles were most frequent and referential bundles least common in university classrooms. Instead, the current results align with more recent studies (e.g., Liu & Chen, 2020a, 2020b; Wang,

2017), which report that referential bundles now predominate in academic lecture corpora. A similar trend has also been observed in interpreting corpora (Li & Halverson, 2020, 2022).

In contrast, discourse-organizing bundles account for approximately 40% of the "LBs" identified in the human-translated corpus. This distribution resembles the register typically found in TED Talks, where such bundles are frequently used to structure discourse and guide the audience through complex topics (Liu & Chen, 2020a, 2020b). As noted by Crawford Camiciottoli and Querol-Julián (2016, p. 319), this rhetorical strategy supports the delivery of specialized scientific content in an accessible and engaging manner, facilitating comprehension across a wide range of disciplines.

Notably, despite the frequent use of VP-based and dependent clause fragments in the human-translated corpus, this did not correspond to a high frequency of stance bundles—features typically associated with conversational registers. This observation suggests that the register of human-translated subtitles may fall somewhere between classroom teaching and conversational discourse. In contrast, the MT-generated translations appear to occupy a position between classroom teaching and textbooks, based on the functional taxonomy of lexical bundles and the four-register framework proposed by Biber et al. (2004).

To explore the distribution of LBs across subcategories, the MT-generated subtitles were found to contain more stance bundles than their human-translated counterparts, particularly within the subcategory of Intention and Prediction. Bundles such as "I am going to", "Next I will explain", "and I will talk about" reflect the lecturer's intent to direct the course flow and orient the audience. In addition, the MT corpus exhibits a higher frequency of referential bundles related to the Specification of Attributes, including examples such as "the surface of the", "the frequency of the", and "application and development of". While structural and functional distributions differ across the two corpora, some of these differences may be attributable to the subject matter of the respective courses. Domain-specific terminology and instructional routines—particularly in mathematically or economically oriented content—likely contribute to the recurrence of certain lexical bundles in both corpora.

Furthermore, a number of similar LBs were identified in both corpora. This observation contributes to addressing the second research question, which examines how the bundles in each corpus compare and how the findings may inform audiovisual translation practices, particularly in the context of future MOOC subtitle production. The analysis revealed overlapping LBs across the two corpora, including eight exact matches and twelve sets of structurally and functionally comparable bundles distributed across different instructional stages. Notable examples of exact matches include "when it comes to", "see you next time", and "going to talk about", which span various structural and functional subcategories. Although the two corpora differ in register—with the human-translated subtitles tending toward an "oral" style and the MT-generated subtitles aligning more with a "literate" register—they nevertheless share a substantial number of lexical bundles, despite differences in production period and subject discipline.

Table 2

Comparison of Similar Bundle Pairs Across Instruction Stages

Machine-Translated	Human-Translated	Structure/Functional Attributes	Instruction Stage
I'm going to introduce	we're going to introduce	VP/DO	Introduction/Preparation
I'm going to talk	we're going to talk	VP/DO	
next I will introduce	next I'm going to	VP/DO	
in this unit, I'll	In this chapter, we	DC/DO	
we'll introduce the	we're going to introduce	VP/DO	
the main function of	the relationship between the	NP/RE	Topic Elaboration
can be used to	it can be used	VP/RE	
on the other hand	so let us take	DC/DO	Transit/Contrast
let's talk about the	let us explore this	VP/DO	Demonstration/Exploration
let's see what	let's explore this	VP/DO	Demonstration/Exploration
I hope you have	I hope you all	VP/SE	Conclusion/Closure
look forward to seeing	looking forward to seeing	VP/SE	Conclusion/Closure

From the analysis of the two corpora, 12 sets of structurally and functionally similar LBs were identified, as presented in Table 2. The columns display LBs from each corpus (MT-generated translation/human translation), along with their corresponding structural and functional attributes across various instructional stages. Notably, verb phrase fragments (VP Frag) and discourse-organizing bundles (DO) account for nearly half of the observed similarities. This pattern suggests that the MOOC lectures in both corpora may share spoken register features that facilitate learner engagement in an online setting. In particular, the presence of discourse organizers in both sets indicates the use of overt signaling devices to introduce topics, mark transitions, or preview key concepts—consistent with Biber and Barbieri's (2007) findings on pedagogical discourse. The distribution of these bundles across instructional phases is also consistent with previous research

(e.g., Cortes & Csomay, 2007; Csomay, 2013), which highlights variation in LB usage depending on pedagogical function. For instance, discourse organizers and referential expressions (RE) tend to appear more frequently in instructional segments, while referential and stance expressions (SE) are more common during lecture openings. Moreover, the appearance of these similar bundles in both corpora—despite the pre-editing process used prior to machine translation—suggests that professional translators may have opted to preserve these forms due to their instructional relevance, rather than revising or omitting them.

Lastly, based on the two research questions explored in this diachronic study, the findings offer two key implications for the translation of MOOC subtitles. First, audiovisual translations of MOOCs may be meaningfully examined within the framework of the academic lecture register, as MOOCs constitute a form of academic discourse adapted for digital delivery. While subtitled video replaces the spontaneity of classroom speech with more scripted forms, this shift primarily highlights how different translation workflows influence the linguistic realization of pedagogical content. In particular, the comparative analysis of machine-translated and human-translated subtitles reveals how each mode of translation may promote differing degrees of simplification, segmentation, and lexical variation, thereby shaping the register of MOOC discourse in distinct ways. Although certain LBs in MT-generated subtitles resemble spoken discourse on the surface—such as directive or predictive expressions—the predominance of referential bundles and condensed phrasing suggests a stronger alignment with a literate register. In contrast, human-translated subtitles tend to preserve more features of spoken discourse, including repetition and interactive transitions, which help maintain the flow of oral delivery. Future research based on expanded corpora of MOOC subtitle translations may further clarify how their register compares with that of other educational formats, such as classroom lectures and TED Talks.

Second, building on the work of Li and Halverson (2020, 2022), the extracted LBs from each corpus could be further examined in relation to their source-language equivalents, particularly with regard to how explicitly or implicitly instructional content is conveyed—a dimension that remains underexplored in audiovisual translation studies. Once the translational relationships between exact matches or similar bundles and their original-language forms are established, these bundles may be particularly well-suited for inclusion in a Translation Memory (TM), as suggested by Grabowski (2018). This is especially relevant for MOOCs that have yet to be translated into English for international audiences. The presence of these bundles in both MT-generated and human-translated subtitles across various disciplines suggests that they may contribute to greater consistency in subtitle register. However, further research is needed to evaluate their effectiveness in TM systems and to assess their potential impact on translator efficiency.

3. Conclusion

This study examined lexical bundles in machine-translated (MT) and human-translated corpora of four MOOCs. Notably, the translated subtitles exhibited a high density of lexical bundles, comparable

to those found in the larger academic corpus analyzed by Biber et al. (2004). Applying their structural and functional taxonomy, the analysis revealed that MT-generated subtitles tend to align with a "literate" register, while human-translated subtitles display more features of an "oral" register. In the MT corpus, referential bundles predominated, whereas discourse organizers accounted for nearly 40% of the bundles in the human-translated corpus. While these results diverge from Biber and Barbieri's (2007) earlier findings, they align with more recent studies suggesting that MT subtitles resemble the style of academic lectures, whereas human-translated subtitles reflect the communicative tone of TED Talks. Despite differences in production periods and disciplinary content, both corpora yielded exact matches and structurally and functionally similar lexical bundles distributed across various instructional stages—highlighting the continuity of academic discourse features in subtitled translation across translation modes.

This study concludes with two key implications. First, MOOC subtitle translation can be productively analyzed within the framework of academic lecture registers. Second, frequently recurring lexical bundles may serve as useful components in Translation Memory (TM) systems for educational subtitling, particularly within audiovisual translation (AVT) contexts.

More broadly, the study contributes to the growing corpus-based AVT research by demonstrating how lexical bundle analysis can reveal systematic register variation across translation modes. As Bruti (2020, 2025) points out, corpus methodologies not only enable the empirical validation of linguistic tendencies in subtitled texts but also provide practical insights into recurring translation strategies and common stylistic choices, offering potential benefits for translator training and quality assurance. In line with Pavesi's (2019) call for more functionally oriented and descriptive AVT studies, this analysis illustrates how corpus-informed approaches, even when applied to smaller datasets, can yield meaningful observations on how translation workflows shape the stylistic and pedagogical tone of MOOC subtitles. Beyond reaffirming the utility of lexical bundles as markers of register, the findings underscore the broader methodological value of corpus linguistics for analyzing instructional, multimodal, and cross-cultural AVT products.

References

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes, 31*(2), 81–92. https://doi.org/10.1016/j.esp.2011.08.004

Anthony, L. (2024). AntConc (Version 4.3.1) [Computer software]. Waseda University. https://www.laurenceanthony.net/software/antconc/

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62*(1), 67–82. https://doi.org/10.1016/j.jml.2009.09.005

Aston, G. (2018). Acquiring the language of interpreters: A corpus-based approach. In M. Russo, C. Bendazzoli, & B. Defrancq (Eds.), *Making way in corpus-based interpreting studies* (pp. 83–96). Springer.

Atapattu, T., & Falkner, K. (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. https://doi.org/10.18608/jla.2018.53.12

Berūkštienė, D. (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. https://doi.org/10.15388/Klbt.2017.11181

Berman, A. (1992). *The experience of the foreign: Culture and translation in romantic Germany*. State University of New York Press.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. John Benjamins.

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., Conrad, S., & Leech, G. (2002). *Longman student grammar of spoken and written English*. Pearson.

Biel, Ł. (2018). Lexical bundles in EU law: The impact of translation process on the patterning of legal language. In S. Goźdź-Roszkowski, & G. Pontrandolfo (Eds.), *Phraseology in legal and institutional settings: A corpus-based interdisciplinary perspective* (pp. 11–26). Routledge.

Boiko, Y. (2023). Diachronic plurality in translation of Shakespeare's plays: A cognitive-discursive perspective. *Cognition, Communication, Discourse*, 26, 41–67.

Bruti, S. (2020). Corpus linguistics and audiovisual translation. In L. Bogucki, M. Deckert, & M. Sokoli (Eds.), *The Palgrave handbook of audiovisual translation and media accessibility* (pp. 381–396). Palgrave Macmillan.

Bruti, S. (2025). Corpora and translation of audiovisual texts. In D. Li, & J. Corbett (Eds.), *The Routledge handbook of corpus translation studies* (pp. 451–465). Routledge.

Butkuvienė, K., & Petrolionė, L. (2023). Diachronic research into translation norms: The case of literary discourse. *Vertimo Studijos*, 16, 29–44.

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

Cortes, V., & Csomay, E. (2007). Lexical bundles in speech and writing. In G. Parodi (Ed.), *Working with Spanish corpora* (pp. 217–230). Continuum.

Coxhead, A., Dang, T. N. Y., & Mukai, S. (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. https://doi.org/10.1016/j.jeap.2017.11.001

Crawford Camiciottoli, B. (2007). *The language of business studies lectures: A corpus-assisted analysis*. John Benjamins.

Crawford Camiciottoli, B., & Querol-Julián, M. (2016). Lecture. In K. Hyland, & P. Shaw (Eds.), *The Routledge handbook of English for academic purposes* (pp. 309–322). Routledge.

Csomay, E. (2013). Lexical bundles in discourse structure: A corpus-based study of classroom discourse. *Applied Linguistics, 34*(3), 369–388. https://doi.org/10.1093/applin/ams045

Das, A., & Das, P. P. (2019). Automatic semantic segmentation and annotation of MOOC lecture videos. In A. Jatowt, A. Maeda, & S. Y. Syn (Eds.), *Digital libraries at the crossroads of digital information for the future* (pp. 181–188). Springer.

Díaz Cintas, J., & Remael, A. (2021). *Subtitling: Concepts and practices*. Routledge.

Ellis, N. C., & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory, 5*(1), 61–78. https://doi.org/10.1515/CLLT.2009.003

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 42*(3), 375–396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

Grabowski, Ł. (2018). On identification of bilingual lexical bundles for translation purposes: The case of an English-Polish comparable corpus of patient information leaflets. In R. Mitkov, J. Monti, G. Corpas Pastor, & V. Seretan (Eds.), *Multi-word units in machine translation and translation technology* (pp. 181–200). John Benjamins.

Hermans, T. (1999). *Translation in systems: Descriptive and systemic approaches explai*ned. Routledge.

Hyland, K. (2008). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41–62.

Hyland, K. (2009). *Academic discourse: English in a global context*. Bloomsbury Publishing.

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.

Lee, C. (2013). Using lexical bundle analysis as discovery tool for corpus-based translation research. *Perspectives: Studies in Translatology, 21*(3), 378–395. https://doi.org/10.1080/0907676X.2012.657655

Lee, T. E. (2024). Exploring naturalness in MOOC video lecture MT subtitles: A case study. *Translogos, 7*(1), 41–61.

Lefevere, A. (1992). *Translation, rewriting and manipulation of literary fame*. Routledge.

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal study. *Modern Foreign Languages, 39*(2), 246–256. https://doi.org/10.1016/j.jslw.2009.02.001

Li, Y., & Halverson, S. L. (2020). A corpus-based exploration into lexical bundles in interpreting. *Across Languages and Cultures, 21*(1), 1–22. https://doi.org/10.1556/084.2020.00001

Li, Y., & Halverson, S. L. (2022). Lexical bundles in formulaic interpreting: A corpus-based descriptive exploration. *Translation and Interpreting Studies, 19*(2), 33–56. https://doi.org/10.1075/tis.19037.li

Liu, C. Y., & Chen, H. J. H. (2020a). Analyzing the function of lexical bundles in undergraduate academic lectures for pedagogical use. *English for Specific Purposes*, 58, 122–137. https://doi.org/10.1016/j.esp.2019.12.003

Liu, C. Y., & Chen, H. J. H. (2020b). Functional variation of lexical bundles in academic lectures and TED talks. *Register Studies, 2*(2), 176–208. https://doi.org/10.1075/rs.18003.liu

Liu, K., & Afzaal, M. (2021). Translator's style through lexical bundles: A corpus-driven analysis of two English translations of Hongloumeng. *Frontiers in Psychology*, 12. https://doi.org/10.3389/fpsyg.2021.633422

Liu, K., Cheung, J. O., & Moratto, R. (2022). Lexical bundles in fictional dialogues of two Hongloumeng translations: A corpus-assisted approach. In R. Moratto, & D. Li (Eds.), *Advances in corpus applications in literary and translation studies* (pp. 229–253). Routledge.

Munday, J. (2012). Introducing translation studies: Theories and applications. Routledge.

Neely, E., & Cortes, V. (2009). A little bit about: Analyzing and teaching lexical bundles in academic lectures. *Language Value, 1*(1), 17–38. https://www.e-revistes.uji.es/languagevalue

Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics, 11*(3), 283–304. https://doi.org/10.1075/ijcl.11.3.04nes

Pavesi, M. (2019). Corpus-based audiovisual translation studies: Ample room for development. In L. Pérez-González (Ed.), *The Routledge handbook of audiovisual translation* (pp. 315–333). Routledge.

Pérez-González L. (2014). *Audiovisual translation: Theories, methods and issues*. Routledge.

Plevoets, K., & Defrancq, B. (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. https://doi.org/10.1075/intp.00001.ple

Pym, A. (2010). *Exploring translation theories*. Routledge.

Salazar, D. (2011). *Lexical bundles in scientific English: A corpus-based study of native and non-native writing* [Unpublished doctoral dissertation]. University of Barcelona.

Shi, J., Otto, C., Hoppe, A., Holtz, P., & Ewerth, R. (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* (pp. 11–19). https://doi.org/10.1145/3347451.3356731

Simpson, R. (2004). Stylistic features of academic speech: The role of formulaic expressions. In U. Connor, & T. A. Upton (Eds.), *Discourse in the professions: Perspectives from corpus linguistics* (pp. 37–64). John Benjamins.

Toury, G. (1995). *Descriptive translation studies—and beyond*. John Benjamins.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning, 61*(2), 569–613. https://doi.org/10.1111/j.1467-9922.2010.00622.x

Uchidiuno, J. O., Ogan, A., Yarzebinski, E., & Hammer, J. (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. https://doi.org/10.1007/s40593-017-0159-7

Venuti, L. (1995). *The translator's invisibility: A history of translation*. Routledge.

Vorobyeva, A. (2018). Language acquisition through massive open online courses (MOOCs), opportunities and restrictions in educational university environment. *XLinguae, 11*(2), 136–146. https://doi.org/10.18355/XL.2018.11.02.11

Wang, Y. (2017). Lexical bundles in spoken academic ELF. *International Journal of Corpus Linguistics, 22*(2), 187–211. https://doi.org/10.1075/ijcl.22.2.02wan

Wang, P.-Y., Chiu, M.-C., & Lee, Y.-T. (2020). Effects of video lecture presentation style and questioning strategy on learning flow experience. *Innovations in Education and Teaching International, 58*(4), 473–483. https://doi.org/10.1080/14703297.2020.1754272

Wang, W.-L., & Li, X.-D. (2017). Translation studies communities in Spain and South Korea: A diachronic comparative study. *Translating and Interpreting Studies, 12*(2), 278–309.

Wood, D. (2010). Lexical clusters in an EAP textbook corpus. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 88–106). Continuum.

Wray, A., & Perkins, M. (2000). The functions of formulaic language: An integrated model. *Language and Communication, 20*(1), 1–28. https://doi.org/10.1016/S0271-5309(99)00015-4

Wu, Y. (2021). Lexical bundles in English EU parliamentary discourse: Variation across interpreted, translated, and spoken registers. *Compilation and Translation Review, 14*(2), 37–86.

Xia, Y. (2014). *Normalization in translation: Corpus-based diachronic research into twentieth-century English-Chinese fictional translation*. Cambridge Scholars Publishing.

Yu, X. (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. https://doi.org/10.1016/j.jeap.2021.101079