5 *Using Corpus Linguistics in Formulaic Language Learning*

Phoebe Lin

5.1 Introduction

Human language can offer infinite creativity, yet it is not always needed for routine social functions.[*] As American linguist Dwight Bolinger (1976: 1) puts it, 'our language does not expect us to build everything starting with lumber, nails, and blueprint, but provides us with an incredibly large number of prefabs, which have the magical property of persisting even when we knock some of them apart and put them together in unpredictable ways'. These prefabricated language chunks are commonly known as *formulaic language*. Wray (2002: 9) coined this term to describe all types of word strings that appear to be 'stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar'. Examples include idioms (e.g., *spill the beans*, *kick the bucket*), proverbs (e.g., *a stitch in time saves nine*), sayings (e.g., *time and tide wait for no man*), cliches (e.g., *better safe than sorry*), speech formulas (e.g., *I'll love you and leave you*), sentence stems (e.g., *it goes without saying*), binomials (e.g., *safe and sound*), and more. Formulaic expressions exhibit a range of properties to varying degrees, such as fixedness in form, semantic non-compositionality, phonological coherence, situation boundedness (as illustrated by the example *without further ado*), and marked grammaticality (as illustrated by the example *by and large*).

In corpus linguistics, Sinclair (1991) emphasised the importance of formulaic expressions in language production, suggesting that language users have a large number of semi-constructed phrases at their disposal. These phrases are used to construct utterances until it becomes necessary to switch to rule-based language. In an analysis of

[*] The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 15600121).

the London Lund corpus by Erman and Warren (2000), these phrases accounted for most of the content of spoken and written texts (58.6 and 52.3 per cent respectively), supporting Sinclair's (1991) arguments.

Despite their importance and prevalence in communication, English as a foreign language (EFL) learners often struggle to learn and memorise formulaic expressions (Lin, 2021). Consequently, these expressions are considered 'the last and most challenging hurdle' in achieving near native-like performance in a second language (Spöttl & McCarthy, 2004: 191). EFL learners also tend to lack a broad knowledge of formulaic expressions (Nguyen & Webb, 2016); make collocation errors (Nesselhauf, 2005); misunderstand idioms in academic lectures (Littlemore et al., 2011); fail to notice formulaic expressions (Martinez & Murphy, 2011); and struggle with rote learning of their forms (Lin, 2021; Wray & Fitzpatrick, 2008).

Corpus linguistics has been striving to help EFL learners master formulaic language. Studies have investigated formulaic language use in spoken and written contexts, learner collocation errors, corpus tool design for facilitating formulaic language learning, the development of formulaic language lists, and so on. These studies are notable for their strong link between theory and practice.

The IdiomsTube project is making direct impacts on EFL teaching, learning, and formulaic language research by applying corpus linguistic methods. This chapter highlights its innovations in corpus compilation and concordancing, its approach to bridging theory and practice in Teaching, and its impacts on pedagogy and research.

The chapter is structured into four main sections following this introduction. Section 5.2 outlines the IdiomsTube project and its theoretical and empirical foundations, focusing on the role of speech prosody in the acquisition of formulaic expressions and the rationale for leveraging video-streaming as pedagogical input. Section 5.3 introduces the IdiomsTube app, describing its core features as a computer-assisted vocabulary learning tool and a web-based multimodal concordancer, alongside the empirical studies that informed its development. Section 5.4 concludes the chapter with a critical reflection on the contributions and innovations of the IdiomsTube project, highlighting its impact on corpus linguistics, pedagogy, and formulaic language research.

5.2 The IdiomsTube Project: An Overview

The IdiomsTube project targets two enduring challenges in English language pedagogy: the struggle of EFL learners to master formulaic

expressions, and the absence of systematic guidance for enhancing vocabulary learning via video-streaming. The goal of this approach is to connect conventional educational strategies with digital technologies, making use of the vast online resources available for language learning.

The project's name, *IdiomsTube*, reflects its focus on the learning opportunities YouTube offers for mastering formulaic language. In the research literature, idioms are a specific type of formulaic expression. They are lexicalised word combinations whose overall meaning does not equal the sum of their parts (e.g., *spill the beans*, *kick the bucket*). While the term *formulaic expressions* more accurately describes the project's full focus, the term idioms was chosen for the project name because it is more familiar to learners and teachers.

5.2.1 Theoretical and Empirical Foundations

The IdiomsTube project is based on the premise that the acquisition of formulaic expressions is fundamentally driven by natural speech prosody (Lin, 2012). It begins when the foetus in the womb hears prosodic cues of the mother's voice and the outside world. Empirical evidence supporting this claim comes from two sources: biological studies that found newborns' cries mimic their first languages' intonation, and linguistics experiments that demonstrate newborns' memory and recognition of tunes that they frequently hear before birth. This biologically predisposed reliance on prosody continues as children grow. Prosodically structured units that are frequent and associated with particular contexts are memorised as holistic units (aka formulaic expressions). These expressions in turn play a key role in first language acquisition.

However, the conditions are very different in the case of EFL learners. These learners lack exposure to diverse communicative and spoken contexts in English outside the classroom. The lack of exposure to prosodic cues in natural speech has made formulaic expression learning in EFL particularly challenge. Recent findings have consistently shown the advantage of spoken input for learning formulaic expressions. In Lin's (2021) controlled experiment, both L1 English and EFL learners showed significantly greater recall of novel multiword combinations presented auditory than textually. This advantage held true even when EFL learners learned novel multiword expressions in discourse contexts (Lin, 2023b).

To help EFL learners, Lin (2012) proposed the broader use of English television and videos in English language teaching to

increase their exposure to diverse spoken communicative contexts. Since then, a series of studies involving corpora of English videos was conducted to explore how online streaming can best benefit formulaic expression learning among EFL learners. These studies investigated various topics, including the authenticity of formulaic expressions on internet television and which genres of internet television may be best for formulaic expressions (Lin, 2014); effective strategies for English vocabulary learning from internet television (Lin & Siyanova-Chanturia, 2014); the prosodic and gestural salience of formulaic expressions in YouTube videos (Lin & Chen, 2020).

A second line of corpus research under the IdiomsTube project focused on describing the prosodic prominence of English formulaic expressions. These studies identified distinctive prosodic patterns in formulaic expressions using multimodal corpora (Lin, 2010, 2018a, 2018b), prosodically annotated corpora (Lin, 2013), and YouTube videos as time-aligned multimodal corpora (Lin, 2019; Lin & Chen, 2020). The findings showed that formulaic expressions often align with intonation units (Lin & Adolphs, 2009; Lin, 2013) and pauses (Wray, 2004). Pitch accents are half as likely to fall on words within formulaic expressions (Lin, 2013, 2018b). For expressions with flexible slots (e.g., *as far as ___ is concerned*; *from the ___ point of view*), the likelihood of pitch accent falling on the words filling these slots seems to correlate with the expressions' frequency in everyday speech. For instance, the flexible slot in *as far as ___ [is] concerned* is far less likely to receive pitch accent when it is filled by *I am* instead of *England football is*. Formulaic expressions are generally less likely to be accented, but there are times when they are prosodically manipulated to achieve emphasis in spontaneous spoken discourse (Lin & Adolphs, 2009; Lin, 2018b). For expressions that can serve multiple discourse functions (e.g., *I think*), their functions may be disambiguated by considering prosodic and gestural cues (Lin & Chen, 2020). These findings underscore the importance of multimodality in formulaic language research.

5.2.2 The IdiomsTube App

The IdiomsTube app began development in 2015, when video-streaming was on the rise. It was a logical next step in a project that built upon the premises of the fundamental role of speech prosody and the need for extensive exposure to verbal input and television in EFL formulaic language acquisition. To address the challenge EFL learners face in mastering formulaic language, we needed an effective way for

learners to gain exposure to natural English speech prosody in diverse communicative contexts. Video-streaming served this purpose, leading to the creation of the IdiomsTube app.

The IdiomsTube app (Lin, 2022b) is the first web-based tool to facilitate formulaic language learning through video-streaming (see Figure 5.1).[1] Prior to this, there were tools to aid the learning of formulaic expressions from *written* input, such as web browser add-ons (Wible, 2008). However, a tool that aids learning from *spoken* input was needed, as it is more conducive to formulaic expression learning than written input (Lin, 2021).

The IdiomsTube app needed a corpus of YouTube videos to work with. The options back then were to compile either a conventional corpus, one that approximates the Movie Corpus, or an instant and transient web corpus made up of captions retrieved in real-time using the YouTube API according to the search term(s) specified by the app user.[2]

The advantage of a conventional and preconstructed corpus of YouTube videos for the IdiomsTube app is that there is ample time

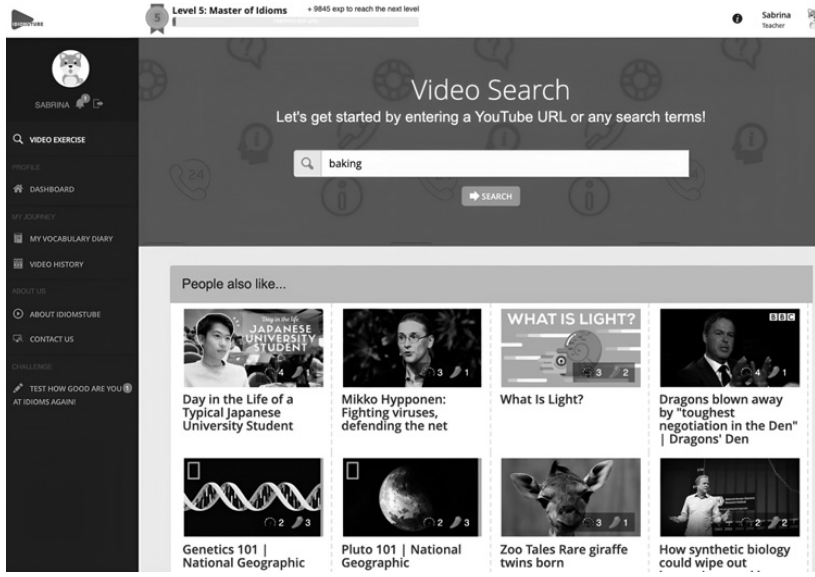

Figure 5.1 The homepage of IdiomsTube with the video search box and recommended videos.

[1] www.IdiomsTube.com. [2] The movie corpus: www.english-corpora.org/movies/.

to design a YouTube corpus that adheres to core values of corpus compilation, including representativeness, balance, and size. Such a corpus will likely be enduring and authoritative. Exercises can also be preconstructed at higher pedagogical quality and larger quantity for each YouTube video. This approach would create an app that delivers predictable and stable performance and involves fewer technical challenges.

The IdiomsTube app chose the latter approach, relying instead on an instant and transient web corpus compiled in real-time based on input from the user. This enhanced both the accessibility and the appeal of the app.

An IdiomsTube app powered by instant and transient web corpora is highly accessible because learners can perform multimodal concordancing with a simple mouse click. This provides a far superior improvement on user experience because previous preconstructed video corpora were purely textual. Users needed to manually locate movie scenes, if they wanted to examine the concordances from multimodal perspectives. As Cheng (2021) demonstrated, in the eyes of learners, the accessibility of a concordancer takes precedence over its capabilities or the quality of the corpus. Learners went to *Google Search* and *NetSpeak* (Potthast et al., 2010) as their go-to tools for error correction, even though traditional concordancers like English-corpora.org (COCA) adhere more strictly to best practices in corpus design, despite having received proper training in using these traditional concordancers.

In the era of Web 2.0, current video trends heavily influence user choices. Due to the search and recommendation algorithms of streaming websites, trending and viral videos receive more traffic than less popular ones, making them more visible, appealing, and in demand among learners. As online trends are transient, corpora and concordancers designed to guide language learning through videos must keep pace with trend changes by instantly compiling video corpora, even if they do not possess qualities such as representativeness, balance, and authority.

YouTube was chosen as the platform for the app because it is the largest source of audio-visual content in English in the Web 2.0 era. As Wojcicki (2015) stated, 400 hours of new content are uploaded to YouTube every minute. Its size and popularity among content creators and users allow it to offer a wide variety of videos, from documentaries and lectures to vlogs, how-to videos, and shows. This diversity allows learners to find videos that match their personal interests and proficiency levels, making learning English through videos both engaging and motivating. Also, streaming on YouTube is straightforward, smooth, and free, providing a positive user experience. Last, the

mature technology behind YouTube, including its application programme interfaces (APIs) and associated unofficial apps, has made it easy to construct new apps around the video-streaming platform.

5.2.3 The IdiomsTube App as a Computer-Assisted Vocabulary Tool

On the IdiomsTube app's front page (see Figure 5.1), users can search for videos currently available on YouTube in English. The app returns fifty videos matching the search terms from the YouTube API. These videos include user-uploaded English captions and contain at least 1,000 words in the caption. Each video's difficulty level is calculated and displayed on the video thumbnail to assist users in choosing videos. Below the search results, the app has sections presenting trending videos among IdiomsTube users, teachers' recommendations (for users enrolled as a class), and users' last-viewed videos.

Before viewing, users are presented a glossary of noteworthy expressions from the video chosen (see Figure 5.2). The glossary includes hyperlinks to the corresponding entries in online dictionaries. The video is then streamed within the IdiomsTube app using the embedded official YouTube player, which allows users to alter playback speed and optionally enable captions. After viewing, users may choose among three types of exercises designed to facilitate recall of the pragmatic use, orthographic, and phonological form of the formulaic expressions.

For example, recall of the pragmatic use of items (see Figure 5.3) is facilitated using a cloze task. IdiomsTube automatically generates cloze exercises to raise learners' awareness of the ways in which formulaic expressions are used in context. Users fill in blanks in the original lines from the video with the expressions. Users can play back corresponding scenes for hints if they are unsure of the correct answers, by clicking the Hint button to return to the exact scene where the expression occurs. The multiple-choice questions are automatically scored; correct answers in the cloze, spelling, and pronunciation tasks increase users' IdiomsTube experience points, raising their ranks within the IdiomsTube world and school-based leaderboards. The app allows users to bookmark expressions for later revision and provides pages for presenting watch history and vocabulary-exposure history. Both history pages include direct links to play back scenes involving each expression (see Figures 5.4 and 5.5). The app also compiles flashcards automatically to help refresh users' memory of all items encountered in previously viewed videos.

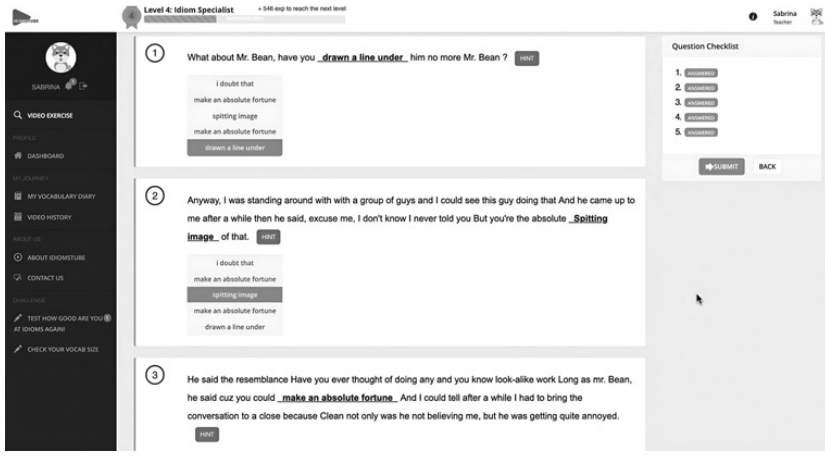

Figure 5.2 A glossary of noteworthy formulaic expressions and hyperlinks to corresponding dictionary entries.

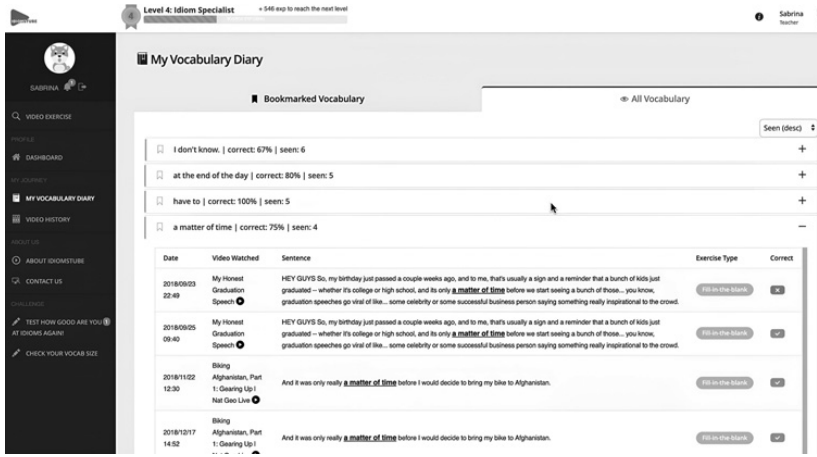

Figure 5.3 The fill-in-the-blank exercise.

In summary, the IdiomsTube app provides EFL learners with three types of support.

1. Noticing and recall. The app highlights noteworthy expressions in the caption and creates learning tasks before, during, and after each video. This structures the English learning process through video-streaming.

Figure 5.4 A summary of expressions learned. Learners can see the context in which each expression was used in any previously watched videos and play back the specific scenes.

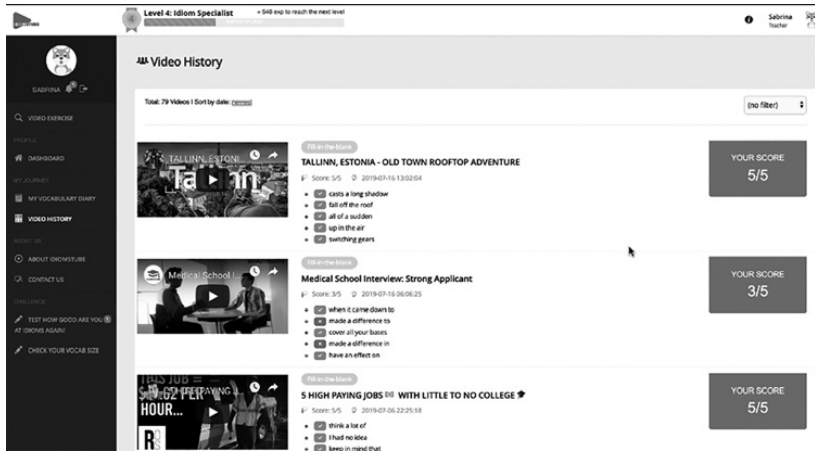

Figure 5.5 The video history page listing expressions present in each video. Learners can play back the specific scenes in which each expression was used in context.

2. The search for comprehensible input. The app evaluates the difficulty of vocabulary and speech rate in videos, enabling learners to select videos that align with their proficiency levels.

3. Motivation and interest. The app suggests videos based on viewing trends among other app users and recommendations from teachers.

5.2.4 The IdiomsTube App as a Concordancer

Concordancing has a rich history in human civilisation. The Cambridge Dictionary defines a concordance as a collection of sentence excerpts containing a specific term. Concordances of the Bible, which help users find related scriptures, are a classic example. This traditional format of concordancing continues to influence its application today in lexicography, corpus linguistics, and language learning. For years, corpus linguists, including myself, have advocated for the use of concordancing in classrooms, arguing that exposure to authentic language use across diverse contexts is crucial for EFL learning (Lin, 2023c, 2026; see also Crosthwaite and Gazmuri Sanhueza, Chapter 2, Rees, Chapter 3, and Apavaloae & Farr, Chapter 4 in this volume). Rather than explicitly teaching vocabulary or grammar rules, concordancing allows EFL learners to observe how meaning is created and situated within context. Reading concordances cultivates cognitive skills vital for language learning, such as predicting, observing, noticing, analysing, interpreting, inferencing, hypothesising, verifying, and more (O'Sullivan, 2007: 277). This process simulates what learners naturally do outside the classroom (Boulton, 2010).

However, the IdiomsTube project raises questions about how to modernise concordancing to make it more relevant and accessible to language learners. More specifically, the design of the IdiomsTube app prompted critical reflections on certain aspects of traditional concordancing.

1. Is the collective display of all instances of a search term the most effective concordancing format for language learners?
2. How much do learners understand about the creation of meaning in context from reading a *textual* concordance line?
3. Should the goal of corpus design be to build a representative and balanced corpus?

While the traditional concordancing approach serves the needs of lexicographers and corpus linguists aiming to uncover language usage patterns, it may not be suitable for learners. First, learners prioritise using the language and broadening their linguistic competence, rather than deeply investigating how a specific lexical or grammatical item is used across contexts. Second, learners may not

have the same level of linguistic knowledge and skills as experts to vividly understand and interpret each *textual* concordance line. Therefore, they may not gain as much from examining a collection of all instances of an item as educators and corpus linguists might expect. Learners may be able to make some observations in the concordance when given prompts and hints from the teacher and working in small groups. However, they might struggle with understanding and interpreting concordances when trying to do so independently outside the classroom, especially without any prompts or hints.

Concordancing typically offers a bird's-eye view of a search term's usage in context by presenting all instances of the term from a specified corpus. This method seems to conflict with standard practices in EFL vocabulary learning, which encourage deep engagement with a text before making inferences about the situated meaning of new vocabulary based on contextual clues. However, the IdiomsTube app, as a concordancer designed for vocabulary learning, bridges these seemingly incompatible methods. It encourages learners to engage deeply with each video before showcasing multimodal concordances of formulaic expressions through a post-viewing cloze exercise. Additionally, it juxtaposes all instances of the expression from previously viewed videos on the vocabulary history page. This approach allows learners to see an overview of the vocabulary usage across different contexts.

While the IdiomsTube app serves as a computer-assisted vocabulary learning tool, it operates as a web-based corpus compiler and concordancer that provides multimodal concordancing on instant web corpora. Using any search term, the app gathers video captions via YouTube's API to instantly compile a time-aligned multimodal web corpus. This corpus undergoes several processes, such as tokenisation, parsing, lemmatisation, speech rate estimation, and lexical difficulty assessment. It also indexes formulaic expressions and their corresponding timestamps in the video caption. These timestamps allow YouTube videos to function as a multimodal corpus, and the IdiomsTube app as a multimodal concordancer. The multimodal concordances are displayed in the cloze exercise, video history page, and vocabulary history page. This allows users to replay scenes and observe the context, gestures, and tone of voice when the formulaic expressions were used. These expressions are identified from a list of 40,000 English formulaic expressions, compiled for the IdiomsTube project (see Section 5.3). These expressions are arranged based on 'learnability', a new measure derived from empirical data. If a video has many formulaic expressions, priority is given to those

with the highest learnability measures. The system currently covers six items, and once a formulaic expression has been shown the number of times specified by the learnability measures (see discussion in Section 5.3.2), the system moves on to the next items based on learnability ranking. Thus, the complexity of the expressions featured in the app increases with the hours of exposure to YouTube videos within the app.

In spoken communication, various elements, including the physical setting, proxemics, gestures, facial expressions, tone of voice, and words, all contribute to creating contextual meanings. However, non-verbal contextual cues are hidden in textual concordancing, which can hinder comprehension and interpretation. The effect of this loss on language learners has yet to be empirically assessed. It is plausible that a learner's ability to imagine these non-verbal cues and understand the context decreases with their linguistic competence, metalinguistic awareness, and exposure to the foreign language. Multimodal concordancing, which provides audio-visual and textual records of spoken discourses, clearly overcomes the limitations of textual concordancing and will be more comprehensible to learners. It was previously unavailable (at least on a large scale) due to technical bottlenecks, including issues related to media storage and redistribution, aligning the transcript with the audio-visual stream, and copyright restrictions.

The IdiomsTube app maximises exposure to multimodal concordancing to help EFL learners understand and learn formulaic expressions. The app offers this through the cloze exercise, video history, and vocabulary history pages in two different formats. The cloze exercise and video history page display multimodal concordances of expressions from individual YouTube videos. The vocabulary history page shows multimodal concordances of previously encountered expressions across videos. These pages are easy to navigate; learners can click a button to view the corresponding time-aligned audio-visual stream. Future enhancements of the multimodal concordancing experience will be available, including the incorporation of in-video motion tracking, automatic on-screen annotations, real-time speak-along with instant pronunciation feedback, and writing practice with AI-powered feedback (Lin, 2023).

Multimodal concordancing could revolutionise concordancing as EFL learners need non-verbal cues to comprehend contextual meanings and remember lexical items. Unlike textual concordance, which only activates memory of orthography, multimodal concordance involves multi-sensory memory, including orthography, phonemics,

tone of voice, hand and facial gestures, proxemics, and more. According to Paivio's (1971) dual coding theory, encoding these additional perceptual memory layers enhances the ease of later retrieval of lexical items.

From a phonological perspective, multimodal concordancing is crucial for learning formulaic language because these expressions have relatively fixed prosodic patterns (Ashby, 2006; Lin, 2012, 2013, 2018). EFL learners need to use the correct accentual pattern for each formulaic expression to accurately convey the intended idiomatic meaning. Otherwise, the expression may be interpreted literally (e.g., **it is raining CATS and dogs*; **my EARS are burning*).

In a nutshell, the IdiomsTube app demonstrates an innovative alternative to traditional concordancing methods for language learners. Like all market-ready products, it prioritises user experience and caters to learners' immediate needs, such as improving their command of formulaic expressions through YouTube videos of their choice. The app pioneers a concordancer with the following characteristics.

1. It is a web-based app that combines a corpus and a concordancer.
2. The corpus is instantly compiled, prioritising timeliness over criteria such as representativeness, balance, or size (see Figure 5.1).
3. The concordance is multimodal (see Figures 5.2, 5.4, and 5.5).
4. Search terms are predetermined for users (see Figure 5.2).
5. Concordancing occurs after a deep engagement with the context and understanding of the text (see Figures 5.3 and 5.5).
6. It does not emphasise presenting multiple instances of the search term.

5.3 Research in the App Development Process

Empirical has played a crucial role in the success of the IdiomsTube app. During the app's development, a series of studies were conducted to inform methodological choices. These studies addressed various practical tasks for the app, such as compiling a comprehensive list of general English idiomatic expressions, ranking these expressions by learnability, and assessing video difficulty for EFL learners. While there is room for improvement in these studies, they have laid a scientific foundation for the IdiomsTube project and opened up new possibilities for formulaic language research.

5.3.1 IdiomsTube's List of 40,000 General English Formulaic Expressions

IdiomsTube identifies formulaic expressions in YouTube video captions based on its own list of general English formulaic expressions. This list forms the core of the app, representing all formulaic expressions detectable in captions and thus, teachable by the app. Theoretically, finding these expressions in video captions should be simple, considering they comprise 58.6 per cent of spoken texts (Erman & Warren, 2000). From this, one would expect to find hundreds of formulaic expressions in a typical thirty-minute video, which usually contains about 3,500 words. However, the actual number of formulaic expressions found in each video will vary depending on factors such as the video genre, topic, quality of English captions, and the speaker's backgrounds and styles. The critical limitation will be the comprehensiveness of the formulaic expression list.

Various formulaic expression lists are available in the literature. These lists focused predominantly on English for academic purposes and tend to be conservative and brief. Examples include Simpson-Vlach and Ellis' (2010) 1,989-item Academic Formulas List, Martinez and Schmitt's (2012) 505-item Phrasal Expressions List, Ackermann and Chen's (2013) 2,468-item Academic Collocation List, and Lei and Liu's (2018) 9,049-item Academic English Collocation List. To construct the Academic Collocation List (ACL), for example, Ackermann and Chen (2013) manually filtered over 130,000 collocations from a 37-million-word corpus of academic texts. Only 1.98 per cent of the original set remained.

These existing studies have provided valuable insights into how IdiomsTube should approach building its own general English formulaic expression list. More specifically, the conventional semi-automatic extraction method, which involves manually filtering a large number of frequently occurring word strings from large corpora, seems unfeasible for two reasons. First, manual filtering is tedious and costly. Assuming it takes two minutes to evaluate each item based on Ackermann and Chen's (2013) criteria, reviewing 130,000 collocations would take over 4,300 work hours. Given the length of the ACL (i.e., 2,468 collocations), it would take approximately 1.74 hours to include a valid collocation. This hampers the scalability of the method. Second, the frequency-based extraction rules used in the literature seem ineffective. That 98.02 per cent of the items have to be discarded during manual filtering indicates poor sensitivity of the automatic extraction rules.

The project resolved to construct its own list of general English formulaic expressions by extracting and consolidating the headwords of seventeen online and print dictionaries of idioms, proverbs, sayings, slang, and famous quotes. This strategy yielded the largest and most credible list of general English formulaic expressions to date, encompassing 40,000 items.[3]

Admittedly, substantial manual processes were required to consolidate the raw list of over 200,000 formulaic expressions, obtained by combining the headwords of all the dictionaries. For example, removing duplicates was a tedious process as they were not exact. Types of duplicates observed include:

- various pronoun variations, for example, *according to your own light*; *according to her own light*; *according to his light*; *according to our own light*;
- particle position, for example, *get out of him*; *get him out of*;
- boundary placement, for example, *you can lead a horse to water*; *you can lead a horse to water, but you cannot make it drink*;
- variation in tense, for example, *make contact with something*; *made contact with something*;
- British/American spellings, for example, *a point of honour*; *a point of honor*;
- use of placeholders, for example, *have nothing to do with*; *have nothing to do with something*; *have nothing to do with someone*;
- use of vague words, for example, *beeline it for*; *beeline it for some place*;
- use of copula *be* and infinitives, for example, *kill two birds with one stone*; *to kill two birds with one stone*; *head over heels*; *be head over heels*.

In addition, entries involving slashes needed manual splitting into two or more formulaic expressions (e.g., *pull/yank one's chain*, *a chill/shiver runs up/down someone's spine* and *there's something/a lot/not much etc. to be said for something*). Disambiguating the headwords into multiple formulaic expressions is too complex to be automated.

The 40,000 shortlisted formulaic expressions were then manually grouped into 'families', a term borrowed from vocabulary research

[3] This list may still not capture the full spectrum of formulaic expressions used in everyday English communication. Future expansion of the list is needed.

literature. For example, the family of *fling one's bonnet over the windmill* includes:

- *fling one's cap over the windmill*;
- *fling one's hat over the windmill*;
- *throw one's bonnet over the windmill*;
- *throw one's cap over the windmill*;
- *throw one's hat over the windmill*.

5.3.2 Determining the Teaching Priority of Formulaic Expressions

Determining the sequence in which new material should be taught is crucial for EFL curriculum design and competence assessment. In vocabulary acquisition research, word frequency has been the basis for deciding this priority. For example, Nation (2012) classified 25,000 English word families into twenty-five bins based on their frequencies in the British National Corpus and Contemporary Corpus of American English. The most frequent 1,000-word families are given priority, and so forth, as these words are more likely to be used in everyday English communication.

However, using frequency to decide the priority of teaching general English formulaic expressions, both within the IdiomsTube app and the broader EFL curriculum, poses challenges. First, general English formulaic expressions lack frequency counts, which means such information needed to be compiled from the ground up for the IdiomsTube project. This task was technically challenging due to the variant forms of formulaic expressions and their typically low frequencies. It is common to find items with the same frequencies, and in such cases, a large number of expressions with the same frequency could be assigned the same learning priority, which would cause confusion.

Second, the decision to prioritise low- or high-frequency formulaic expressions in the IdiomsTube app is contentious. If the app prioritises high-frequency expressions, such as *you know what*, *come on in*, *have a wonderful day*, it might not seem useful or interesting to its current and target users, who are intermediate EFL learners at high school or university levels. These learners are expected to be familiar with these expressions already. On the other hand, focusing on low-frequency expressions presents a dilemma as the chance that users will come across the same expressions in their next chosen video will be extremely low. Given the importance of spaced repetitions in vocabulary learning (Ellis,

1995), the lack of opportunity to re-encounter new vocabulary soon after learning it causes attrition.

To overcome these issues, IdiomsTube developed a new approach to determine the sequence in which formulaic expressions should be taught. This method prioritises the teaching of expressions with the highest level of learnability.

In this project, *learnability* is defined as the number of exposures EFL learners need to recall the form of a formulaic expression. The research to determine the learnability of all 40,000 formulaic expressions involves three parts. The first part includes compiling five norms for all the expressions through corpus analysis and crowdsourcing. These norms include frequency of occurrence in spoken and written corpora, semantic transparency, imageability, and the presence of word association.

The second part is a learning experiment where EFL learners were asked to learn formulaic expressions sampled from the list of 40,000 items. The mean number of repetitions needed to correctly recall the form of 525 formulaic expressions was recorded.

In the third and final part, the response variable from the learning experiment was analysed using a linear mixed-effects (LME) model with the five norms from part one as the fixed effects and subject and item as random effects. The LME model was used to estimate the learnability of the remaining items in the list of 40,000 general English formulaic expressions.

The results of this three-part study include a list of 40,000 general English formulaic expressions, arranged according to learnability, an experimentally derived measure that reflects the estimated number of repetitions needed to learn the expressions, and a database providing norms for five features of the English formulaic expressions. These norms are fundamental to the design of stimuli and materials in future linguistic experiments involving formulaic expressions.

Beyond the IdiomsTube project, the list of English formulaic expressions informs not only EFL vocabulary syllabi but also vocabulary competence tests for clinical populations. For instance, difficulty understanding idiomatic and non-literal language is a key characteristic of developmental language disorders (Benjamin, Newton, & Ebbels, 2020), autism spectrum disorders (Whyte, Nelson, & Scherf, 2014; Saban-Bezalel & Mashal, 2019), and schizophrenia (Titone, Holzman, & Levy, 2002). The list of general English formulaic expressions has been developed into a pilot test for assessing the presence or severity of impairment in non-literal language comprehension (Lin, 2022a). When completed, such

a groundbreaking test would be useful to speech therapists, neurologists, and psychiatrists.

5.4 Critical Reflections on Application and Impact

Corpus linguistics has always played a pivotal role in formulaic language research and pedagogy. The field gained attention when Sinclair (1991) identified the prevalence of formulaic expressions in everyday communication, proposing the Idiom Principle. From the beginning, corpus linguistic methods have been applied to develop solutions for mastering formulaic expressions. Studies have created dictionaries and lists of English formulaic expressions for specific purposes, investigated their coverage in EFL textbooks, identified common usage mistakes among EFL learners, and developed web-based tools to facilitate formulaic language learning, and so on.

This chapter presented the IdiomsTube project, which applied corpus linguistic methods to address gaps in formulaic language research and pedagogy. The project produced:

1. an extensive list of general English formulaic expressions;
2. insights into their unique prosodic patterns;
3. an objective assessment of online television's potential for learning formulaic expressions;
4. a database presenting the frequency, semantic transparency, and imageability norms for 40,000 English formulaic expressions;
5. a scientific basis for deciding the sequence in which formulaic expressions may be taught;
6. a method for assessing a video's difficulty for EFL learners;
7. a web-based app that guides independent learning of formulaic expressions through online streaming; and
8. a test of formulaic language competence for the clinical population.

The IdiomsTube project expanded the scope of formulaic language research and corpus linguistics, making impacts in English language teaching, EFL curriculum development, computer-assisted language learning, and competence assessment in clinical contexts.

The most direct impacts of the project are seen among EFL learners and teachers. The IdiomsTube app, designed to facilitate independent learning of formulaic expressions through YouTube videos, has been popular since its launch in December 2018, attracting 10,000 registered users worldwide. The app incorporates empirical findings about effective formulaic language learning and prioritises user experience, immediate learner needs, and tool accessibility.

To engage teachers and learners, biannual workshops or talks were conducted. A public talk in January 2022, 'What are formulaic expressions and how do we master them?', attracted 300 teachers, researchers, and English learners worldwide. The project's Instagram and Facebook have also connected users and sustained their interest in learning English formulaic expressions through video-streaming. In 2017–2020, new posts were released daily, attracting over 2,000 subscribers and over 198,400 reads.

The IdiomsTube teachers' interface, a dashboard designed for teachers to enrol their entire class in the IdiomsTube app and monitor their progress, has been particularly useful during the COVID-19 pandemic, where many schools used it as daily homework to enhance students' English exposure outside the virtual classroom.

The project's basic research studies have equally gained enthusiastic support from academic researchers, research councils, and education policymakers. As the government organisation in charge of the local English curriculum, the Hong Kong Education Bureau funded the development of the IdiomsTube app for four years and provided many opportunities for the project to present its findings and outcomes. The collaboration was fruitful as formulaic language learning and technology in education are strategic areas of local education policy.

In conclusion, the IdiomsTube project demonstrated many fruitful ways in which corpus linguistics has advanced formulaic language research and pedagogy. At the same time, the project innovates in terms of modernising approaches to corpus and concordancing, such as highlighting the value of instant web corpora and enabling multimodal concordancing on a large scale. More innovations can be expected from the IdiomsTube app as the project incorporates generative AI to offer new functions to further enhance the user experience.

References

Ackermann, K. & Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL): A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*(4), 235–247.

Adolphs, S. & Lin, P. (2011). Corpus linguistics. In Li, W., Zhu, H., & Simpson, J. (eds.), *The Routledge Handbook of Applied Linguistics*. Routledge (pp. 597–610).

Ashby, M. (2006). Prosody and idioms in English. *Journal of Pragmatics*, *38*(10), 1580–1597.

Benjamin, L., Newton, C., & Ebbels, S. (2020). Investigating the effectiveness of idiom intervention for 9–16-year-olds with developmental language disorder. *International Journal of Language & Communication Disorders*, *55*(2), 266–286.
Bolinger, D. (1976). Meaning and memory. *Forum Linguisticum*, *1*, 1–14.
Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language Learning*, *60*(3), 534–572.
Cheng, Y.-H. (2021). EFL college students' concordancing for error correction. *English Teaching & Learning*, *45*(4), 431–460.
Ellis, N. C. (1995). Psychology of foreign language vocabulary acquisition: Implications for CALL. *Computer Assisted Language Learning*, *8*(2), 103–112.
Erman, B. & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, *20*(1), 29–62.
Lei, L. & Liu, D. (2018). The academic English collocation list: A corpus-driven study. *International Journal of Corpus Linguistics*, *23*(2), 216–243.
Lin, P. (2010). The phonology of formulaic sequences: A review. In Wood, D. (ed.), *Perspectives on Formulaic Language: Acquisition and Communication*. Continuum (pp. 174–193).
Lin, P. (2012). Sound evidence: The missing piece of the jigsaw in formulaic language research. *Applied Linguistics*, *33*(3), 342–347.
Lin, P. (2013). The prosody of idiomatic expressions in the IBM/Lancaster Spoken English Corpus. *International Journal of Corpus Linguistics*, *18*(4), 561–588.
Lin, P. (2014). Investigating the validity of internet television as a resource for acquiring L2 formulaic sequences. *System*, *42*(1), 164–176.
Lin, P. (2018a). Formulaic language and speech prosody. In Siyanova-Chanturia, A. & Pellicer-Sanchez, A. (eds.), *Understanding Formulaic Language: A Second Language Acquisition Perspective*. Routledge (pp. 78–94).
Lin, P. (2018b). *The Prosody of Formulaic Sequences: A Corpus and Discourse Approach*. Continuum.
Lin, P. (2019). Self-directed English vocabulary learning through YouTube videos. ELTU Conference 2019, Chinese University of Hong Kong, Hong Kong, 27–28 May.
Lin, P. (2021). In search of the optimal mode of input for the acquisition of formulaic expressions. *TESOL Quarterly*, *55*(3), 1011–1023.
Lin, P. (2022a). Developing a test of comprehension of non-literal language in medical context. Paper presented at the 20th International and Interdisciplinary Conference on Communication, Medicine and Ethics (COMET 2022), the Hong Kong Polytechnic University, 13–15 July 2022.
Lin, P. (2022b). Developing an intelligent tool for computer-assisted formulaic language learning from YouTube videos. *ReCALL*, *34*(2), 185–200.
Lin, P. (2023a). ChatGPT: Friend or foe (to corpus linguists)? *Applied Corpus Linguistics*, *3*(3), 1–10.
Lin, P. (2023b). Computer-assisted learning of English formulaic expressions from YouTube videos. In Reynolds, B. L. (ed.), *Vocabulary Learning in the Wild*. Springer (pp. 309–333).
Lin, P. (2023c). Why corpus linguistics matters to L2 teaching and learning. In Sands, K., Petray, M. J., Clements, G. D. & Santelmann, L. (eds.), *Linguistic*

Foundations for Second Language Teaching and Learning. Cambridge University Press (pp. 1–15).
Lin, P. & Chen, Y. (2020). Multimodality I: Speech prosody and gesture. In Adolphs, S. & Knight, D. (eds.), *Routledge Handbook of English Language and Digital Humanities*. Routledge (pp. 66–84).
Lin, P. & Siyanova-Chanturia, A. (2014). Internet television for L2 vocabulary learning. In Nunan, D. & Richards, J. C. (eds.), *Language Learning Beyond the Classroom*. Routledge (pp. 149–158).
Lin, P. & Adolphs, S. (2009). Sound evidence: Phraseological units in spoken corpora. In Barfield, A. & Gyllstad, H. (eds.), *Researching Collocations in Another Language: Multiple Interpretations*. Palgrave Macmillan (pp. 34–48).
Lin, P. & Adolphs, S. (2024). Corpus linguistics. In Li, W., Zhu, H., & Simpson, J. (eds.), *The Routledge Handbook of Applied Linguistics*, 2nd ed. Routledge (pp. 296–308).
Lin, P. (2026). Corpus Linguistics and Human-Computer Interaction (HCI). In H. Nesi & P. Milin (Eds.), International Encyclopedia of Language and Linguistics (3rd ed.). Elsevier.
Littlemore, J., Chen, P. T., Koester, A., & Barnden, J. (2011). Difficulties in metaphor comprehension faced by international students whose first language is not English. *Applied Linguistics*, *32*(4), 408–429.
Martinez, R. & Murphy, V. A. (2011). Effect of frequency and idiomaticity on second language reading comprehension. *TESOL Quarterly*, *45*(2), 267–290.
Nation, P. (2012). The BNC/COCA word family lists 25,000. Retrieved 10 May 2025 from www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs.
Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. John Benjamins.
Nguyen, T. M. H. & Webb, S. (2016). Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research*, *21*(3), 298–320.
O'Sullivan, Í. (2007). Enhancing a process-oriented approach to literacy and language learning: The role of corpus consultation literacy. *ReCALL*, *19*(3), 269–286.
Paivio, A. (1971). *Imagery and Verbal Processes*. Holt, Rinehart & Winston.
Potthast, M., Trenkmann, M., & Stein, B. (2010). Netspeak: Assisting writers in choosing words. In Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., & van Rijsbergen, K. (eds.), *Advances in Information Retrieval*. Springer (pp. 672–672).
Saban-Bezalel, R., Dolfin, D., Laor, N., & Mashal, N. (2019). Irony comprehension and mentalizing ability in children with and without autism spectrum disorder. *Research in Autism Spectrum Disorders*, *58*, 30–38.
Simpson-Vlach, R. & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*(4), 487–512.
Sinclair, J. (1991). *Corpus, Concordance and Collocation*. Oxford University Press.
Spöttl, C. & McCarthy, M. (2004). Comparing knowledge of formulaic sequences across L1, L2, L3, and L4. In Schmitt, N. (ed.), *Formulaic Sequences: Acquisition, Processing and Use*. John Benjamins (pp. 191–225).

Titone, D., Holzman, P. S., & Levy, D. L. (2002). Idiom processing in schizophrenia: Literal implausibility saves the day for idiom priming. *Journal of Abnormal Psychology*, *111*(2), 313.

Whyte, E. M., Nelson, K. E., & Scherf, K. S. (2014). Idiom, syntax, and advanced theory of mind abilities in children with autism spectrum disorders. *Journal of Speech, Language & Hearing Research*, *57*, 120–130.

Wible, D. (2008). Multiword expressions and the digital turn. In Meunier, F. & Granger, S. (eds.), *Phraseology in Foreign Language Learning and Teaching*. John Benjamins (pp. 163–181).

Wojcicki, S. (2015) Keynote speech. VidCon 2015 Conference, Anaheim Convention Center, California, 23–25 July.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge University Press.

Wray, A. (2004). 'Here's one I prepared earlier': Formulaic language learning on television. In Schmitt, N. (ed.), *Formulaic Sequences: Acquisition, Processing and Use*. John Benjamins (pp. 249–268).

Wray, A. & Fitzpatrick, T. (2008). Why can't you just leave it alone. In Meunier, F. & Granger, S. (eds.), *Phraseology in Foreign Language Learning and Teaching*. John Benjamins (pp. 123–147).

Wynne, M. (ed.). (2005). *Developing Linguistic Corpora: A Guide to Good Practice*. Oxbow Books.