Dai, Y., & Wu, Z. (2026). Learning (semi-)technical music words with mobile dictionary and/or machine-translation apps: performances and perceptions. Computer Assisted Language Learning, 39(3), 593–614. https://doi.org/10.1080/09588221.2024.2412788

This is an Accepted Manuscript of an article published by Taylor & Francis in Computer Assisted Language Learning on 09 Oct 2024 (published online), available at: https://doi.org/10.1080/09588221.2024.2412788.

Learning (semi-)technical music words with mobile dictionary and/or machine-translation apps: Performances and Perceptions

Abstract:

This article examines the use of mobile dictionary and/or machine-translation (MT) apps to facilitate the learning of (semi-)technical music words. A total of 93 Chinese students were randomly assigned into three groups (31 each): the Dictionary Group, the MT Group, and the Dictionary+MT Group. They translated sentences containing the target semi-technical or technical words with the help of mobile dictionary or MT apps or both. It was found that the Dictionary Group had the best retention rates of semi-technical and technical words; the Dictionary+MT Group was in-between; and the MT Group had the worst retention rates. Analysis of dictionary look-up behaviors further showed that the Dictionary Group looked up twice as many (semi-)technical words as the Dictionary+MT Group. These findings are explained from the perspective of Cognitive Load Theory to show how the Dictionary Group had the best level of mental resources germane to vocabulary learning. Despite the inter-group differences of retention rates, perception surveys showed that the groups did not differ in their attitudes towards the dictionary or MT apps, suggesting that student performances and perceptions were misaligned. Based on the research findings, the article offers pedagogical implications for mobile-assisted learning of specialized words.

Keywords: technical words, semi-technical words, dictionary, machine translation, mobile-assisted language learning

1. Introduction

Language students of the 21st century have a wide range of digital resources at their disposal. Two widely used resources are dictionaries and machine translation (MT). Their popularity has been corroborated by multiple survey studies (e.g., Levy & Steel, 2015; Niño, 2009; O'Neill, 2019a). However, these tools are variably perceived by students and teachers. While students are largely receptive to both tools (O'Neill, 2019a), teachers believe that MT may produce poor output and seldom use it in language classrooms (Niño, 2009). By contrast, dictionaries garner more trust from teachers and systematic training has been advocated to improve students' consultation skills to maximize the learning effect (Webb & Nation, 2017). Although teachers tend to favor dictionaries over MT, students may still turn to dictionaries and/or MT for help (Lee, 2023; Peters & Fernández, 2013) when they encounter unknown (specialized) words. These two resources are ever more prevalent with the advent of mobile apps that promise immediate, easy, and direct solutions to lexical problems. Despite their widespread use, we have yet to understand the extent to which dictionary and MT apps can benefit the learning of specialized words. The current study attempts to address this gap by examining how different mobile-assisted conditions lead to the learning of music words, and by extension, specialized words in other fields.

2. Research background

2.1 (Semi-)technical words

Specialized words can be those shared by several fields or specific to one field (Nation, 2022). In this study, we focus on the latter, namely specialized words in the music field. We also distinguish two types of specialized words: semi-technical and technical. Semi-technical words refer to those having both unspecialized and specialized meanings (Smith & Davis, 2018), such as *bar* that can mean a place serving drinks or a smaller section of a piece of music. By contrast, technical words contain specialized meanings only, such as *vibrato*. Learning specialized meanings of semi-technical words is difficult due to their polysemous and opaque nature (Watson Todd, 2017). Relying on the familiar form-meaning mapping, students may mistake the non-technical meaning for the technical meaning (Deignan & Love, 2021). Relatedly, technical words are difficult to learn for two possible reasons. First, the form and meaning of a technical word are both new to the students, who may focus on one at the expense of the other (Haynes & Baker, 1993). Second, technical words tend to be low-frequency words (compared to semi-technical words), which has been found as a predictor of difficulty (Cervetti et al., 2015).

While studies on specialized words abound, the majority have endeavored to produce English-for-specific-purposes (ESP) wordlists, with relatively less attention to the actual learning of (semi-)technical words. As a notable exception, Gablasova's study (2015) revealed that students achieved better results in acquiring specialized words when exposed to their first language than their second language. More recently, Yüksel et al. (2022) found that the Quizlet app (functioning as a digital flashcard) was more effective than wordlists in facilitating technical vocabulary learning. These encouraging results point to the remarkable potential of mobile-assisted learning of (semi-)technical vocabulary. Indeed, a meta-analysis has shown that the effect of mobile-assisted learning for general-purpose vocabulary is large and positive (Lin & Lin, 2019), but the studies included for the meta-analysis have centered on the use of text/multimedia messages, or purposefully-designed vocabulary learning apps. As such, Lin and Lin (2019) rightly pointed out that "few studies focused on the effectiveness of mobile dictionary applications for L2 word retention" (p. 911). We concur with this observation and further contend that much remains unknown as to whether and how mobile dictionary apps can benefit the learning of specialized vocabulary.

2.2 Dictionaries

Dictionary look-up involves multiple levels of cognitive engagement, such as noticing, attending to, and processing word features, all of which facilitate vocabulary learning (Webb & Nation, 2017). However, it should be pointed out that the working memory capacity of a student is limited and how mental effort is distributed will lead to different learning outcomes (Barcroft, 2009). In this regard, Cognitive Load Theory (Sweller et al., 2011) is helpful to explain the relation between dictionary look-up and vocabulary retention (Dai et al., 2019). Three types of cognitive load are distinguished: intrinsic, extraneous, and germane. Intrinsic cognitive load is determined by the inherent difficulty of the learning materials, while extraneous cognitive load is imposed by suboptimal instructional design or information presentation. Germane cognitive load refers to "the working memory resources that the learner devotes to dealing with the intrinsic cognitive load associated with the information" (Sweller,

2010, p. 126). Given the limited capacity of working memory, it is important to reduce extraneous cognitive load so that more mental resources (germane cognitive load) can be freed to handle intrinsic cognitive load. For example, Liu et al. (2014) found that students working with a key-in dictionary achieved better retention of the spellings than those working with a click-on dictionary and those without access to a dictionary. This was because germane cognitive load was sufficiently allocated to the spellings in the key-in condition. Also framed from the perspective of Cognitive Load Theory (CLT), Dziemianko (2022) compared four types of illustrations in digital dictionaries: colorful illustrations, greyscale illustrations, line drawings, and illustration-free texts. She found that line drawings led to the best vocabulary retention because the "economical form" of drawings, without distracting details, could reduce the students' extraneous cognitive load and allow them to focus on the information germane to vocabulary learning (p. 227).

These studies have shown that CLT offers a powerful lens to analyze student engagement with different (digital) tools and it is particularly pertinent to the context of mobile-assisted learning of specialized vocabulary. For example, when students look up specialized words, does a dictionary app induce extraneous or germane cognitive load? A paper dictionary may incur inefficient searches that compete for the limited capacity of working memory (Sweller et al., 2011; Dai et al., 2019). By comparison, an electronic dictionary may require less mental effort to locate a word and potentially reduce extraneous cognitive load (Dziemianko, 2018). In addition, when presented with unknown specialized words, it is not uncommon for students to resort to machine translation (MT) and then crosscheck with dictionary apps. Do these dual steps contribute to germane cognitive load (i.e., more mental resources devoted to learning) or incur extraneous cognitive load (i.e., cognitively distracted by using an additional tool)? The answers to these questions will enable us to understand whether and how these popular apps contribute to the learning of specialized words.

2.3 Machine translation

A systematic review of the use of MT in foreign language learning shows that the effect is generally positive (Lee, 2023). Research has found that students working with MT texts can be sensitized to lexical, grammatical, and syntactic features, thus contributing to their metalinguistic awareness (Niño, 2008). Specific to language skills, writing has drawn substantial scholarly attention but the findings are mixed. One common research design is to compare students' compositions produced with and without the assistance of MT. Studies have revealed that students with access to MT produced more words (Cancino & Panes, 2021; Chon et al., 2021; Garcia & Pena, 2011) and achieved better quality, as indicated by overall writing scores (Garcia & Pena, 2011) and specific textual measures, such as accuracy (Cancino & Panes, 2021), lexical diversity (Fredholm, 2019), lexical complexity (Chon et al., 2021) and syntactic complexity (Cancino & Panes, 2021; Chon et al., 2021). By contrast, Chung and Ahn (2022) observed that the use of MT decreased lexical/syntactic complexity and did not produce longer writing, thus complicating the findings about the learning effects of MT. While these studies have touched upon vocabulary use, little is known about the effect of MT on vocabulary retention. Lo's (2023) study is a rare exception that has examined vocabulary learning assisted by MT. It was found that the use of MT could benefit students' immediate vocabulary learning, but higher-proficiency students achieved better vocabulary retention in the delayed posttest.

Another strand of studies (albeit small in numbers) compared the learning effects of MT and dictionaries. Fredholm (2019) found that MT enabled students to produce compositions with higher lexical diversity than a paper dictionary. Similarly, O'Neill (2019b) compared the compositions produced with the assistance of MT or online dictionaries or without any access to tools. Highest writing scores were observed in the groups using MT. However, both studies showed that when students no longer had access to MT in the (delayed) posttest, no developmental effects were found. These results are contradictory to those reported by Lo (2024), who found that consulting online dictionary and MT had similar developmental effects on vocabulary learning. Given the conflicting results related to online dictionaries and MT, we are interested to know whether MT has an advantage over dictionaries in the mobile-assisted learning of specialized words, and if so, whether the advantage can sustain over time.

2.4 Research gaps and research questions

Based on the studies reviewed previously, three gaps can be discerned and summarized as follows: (a) learning effects, (b) look-up behaviors, and (c) student perceptions. First, compared to the large amount of scholarly effort in producing ESP wordlists, less attention has been paid to the actual learning of specialized words, particularly in MT-assisted contexts. This gap is striking because students commonly use MT to solve their lexical problems (O'Neill, 2019a), and students also perceive vocabulary learning as one of the top-rated benefits associated with MT (Almusharraf & Bailey, 2023). However, O'Neill's (2019b) and Fredholm's (2019) studies revealed that the benefits of MT disappeared when the tool was no longer available to the students; and Klimova et al. (2023) has contended that MT will not contribute to vocabulary retention (but see Lo, 2023, 2024 for a different view). These negative observations stand in contrast to a substantial body of research demonstrating the positive effect of dictionaries on vocabulary learning (e.g., Nation, 2022; Webb & Nation, 2017). Given the popularity of MT and dictionary apps, it is necessary to ascertain the learning effects of these apps so that we will understand whether MT apps are more, less, or equally effective as compared to dictionary apps in facilitating specialized vocabulary learning.

Second, with a limited number of studies directly comparing MT and dictionaries, we know little about how students use these tools in different conditions. Previous studies have compared the use of two individual tools (Fredholm, 2019; Lo, 2024; O'Neill, 2019b), but with mobile phones, students can easily switch between apps. Thus, one pertinent question about mobile-assisted specialized vocabulary learning is whether MT output will be followed up by dictionary look-up. Put differently, when students working with both MT and dictionary apps, do they look up a larger, smaller, or same number of words than those working with the dictionary app only? As dictionary look-up involves multiple facets of cognitive engagement (see Section 2.2), the comparison will provide suggestive evidence to understand students' (in)action on cognitive affordances triggered by MT output. Synthesized with the learning outcomes, it will allow us to know whether (in)access to MT output influences students' dictionary look-up behaviors that potentially benefit or inhibit specialized vocabulary learning.

Third, while previous studies have compared student perception of MT and online dictionaries (O'Neill, 2019a), they have focused on the *online* versions of the tools for *English-for-general-purposes (EGP)* language learning. Much remains unknown about student perception of using *mobile* apps for learning *ESP* words. Given the increasing ubiquity of

mobile apps, we need to understand how students perceive MT and dictionary apps, the two most common tools, in the development of their specialized vocabulary knowledge. The triangulated insights about learning effects, look-up behaviors, and student perceptions will enable instructors to offer incisive advice for students to strategically choose and use digital tools that will benefit their ESP learning. As such, we conducted a study comparing three mobile-assisted conditions for learning music words: dictionary app, MT app, and dictionary+MT apps. The study was guided by the following research questions (RQ):

RQ1: To what extent does the retention of (semi-)technical music words differ among the Dictionary group, the MT group, and the Dictionary+MT group?

RQ2: How does the presence of MT texts influence the students' dictionary look-up behaviors?

RQ3: How do the students perceive using the dictionary and/or the MT apps to learn music words?

3. The study

3.1 Context and participants

This study involved 93 music-majored undergraduate students in China. They had a relatively low level of English proficiency (roughly equivalent to the A2 level in CEFR). At the time of the study, they were enrolled in a third-semester English enhancement course with one of the objectives to expand the students' specialized vocabulary. They were randomly assigned into three mobile-assisted learning conditions (31 students each): the Dictionary (Dic) Group used a dictionary app only; the Dic+MT Group used both dictionary and MT apps; and the MT Group used MT apps only. The students gave their informed consent to participate in the study, which was conducted in compliance with relevant ethical requirements.

3.2 Target words and the translation task

We selected 10 semi-technical words and 10 technical words as the target words (see Appendix A). The semi-technical words had at least one unspecialized meaning and one specialized meaning. The technical words had specialized meanings only. Dang et al. (2022) suggested that words unknown to over 80% of the participants could be included for a study. In a pretest (sentence translation, detailed below), the specialized meanings of these 20 words were unknown to over 90% of the participants in each of the three groups and thus deemed suitable for the current study.

A sentence translation task was adopted in this study based on two grounds. First, translation tasks are useful for vocabulary learning and immediately compatible with a mobile-assisted learning environment (Slatyer & Forget, 2020) where students would feel comfortable translating sentences facilitated by mobile apps. Second, sentence translation tasks are considered to have optimal "reliability and discriminability…most suitable for assessing knowledge of words with multiple meanings" (Lu et al., 2020, p. 316). Applied to our context, the sentence translation task was particularly helpful to assess the students' knowledge of semi-technical music words.

We designed two sets of sentences (Set A and Set B) for the English-to-Chinese translation task, whereby the participants translated the entire sentences. Each set had 20

sentences and each sentence contained only one un-highlighted, target word. The overall difficulty of the two sets was considered comparable because the percentage of high-frequency words (first 3,000 word families) plus common proper nouns (e.g., *Tom* and *Susan*) was the same (94.2%) for both sets. Set A was used for the learning task during class time and also for the pretest and the immediate posttest, while Set B was used for the delayed posttest. Although it might have been ideal to have three sets of sentences for three tests to minimize the practice effect, we decided to have two sets to balance between ensuring the experimental validity and reducing the practice effect. If we had used a different set of sentences for the immediate posttest, students might figure out what were the target words (that were kept constant between the tests), thus compromising the experimental validity. Additionally, using the same set of sentences for the immediate posttest and another comparable set for the delayed posttest could elicit vocabulary learning respectively in the identical (non-transfer) sentential contexts and the similar-yet-different (near-transfer) sentential contexts (see Barenberg et al., 2020 for the concepts of non-transfer and near-transfer).

3.3. Mobile apps

Two types of apps were used in the study: dictionary and MT. A mobile dictionary app, called "Eudic", was used because it allowed the customization of lexicographical databases. We created a database with 60 entries, corresponding to 60 words in the sentences of Set A.

- For the 30 *music* words (10 target semi-technical words, 10 target technical words, and another 10 non-target music words), the lexicographical information included: spelling, pronunciation, parts of speech, bilingual definitions of the *specialized* sense, and bilingual sample sentences. These entries were tantamount to those in an ESP dictionary.
- For the 30 *non-music* words, the app provided the following lexicographical information (where applicable): spelling, pronunciation, parts of speech, bilingual definitions of *all* senses, bilingual sample sentences, and collocations, i.e., exactly the same content one would find in an EGP dictionary.

These design features (mixing music and non-music words, and general-purpose and specialized information) were motivated by three reasons. First, the students could not guess which words were the target words, thus ensuring the experimental validity. Second, they could look up the non-target words (if need be) when they translated the music sentences, thus improving the ecological validity. Third, only specialized information was provided for the music words to allow the low-proficiency students "quick and easy access to" the relevant specialized information during the translation task (Tarp, 2010, p. 41). Framed from a CLT perspective, this design could potentially reduce their extraneous cognitive load arising from unnecessary searches or information overload.

For the MT apps, the students were asked to use the one they were most familiar with. We did not control for the use of MT apps for two reasons. First, if the students had been asked to use a MT app they were not familiar with, then the Dic+MT Group would have been exposed to two new apps (dictionary and MT). This would increase the difficulty of learning new tools and discourage them from using both tools. Second, allowing the students to consult the MT apps they habitually used could increase the ecological validity, without diminishing the experimental validity. This was because the Dic+MT Group and the MT Group reported using

either Youdao or Baidu[1] as their choice of the most familiar MT app, and the frequency distribution did not differ between the two groups (see Appendix B for the Chi-squared test result).

3.4 Procedures

The study was conducted over three weeks (see Figure 1). In the first week, the students took the pretest without the assistance of any tool.

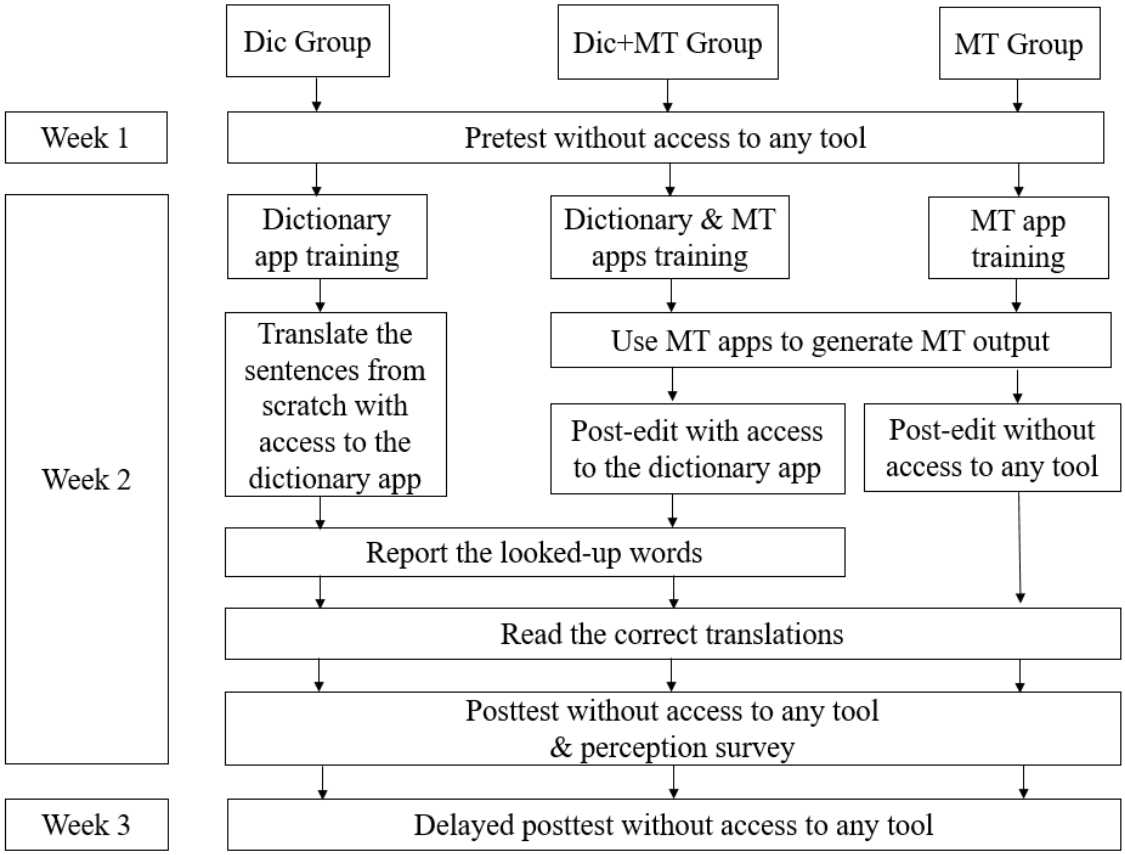

Figure 1. Experimental procedures

In the second week, the students were trained to use the dictionary and/or MT apps effectively. Adapted from O'Neill (2019b), the training was structured in five steps: (a) the instructor explained the functions; (b) the instructor demonstrated how to use the app(s); (c) the students practiced using the app(s); (d) the instructor reviewed the app features; and (e) the instructor answered the students' queries. After the training, the Dic+MT Group and the MT Group first used the MT apps to generate machine-translated output. Then, the Dic+MT Group post-edited the texts assisted by the dictionary app, while the MT Group post-edited the texts without access to any tool. The Dic Group translated the sentences from scratch, facilitated by the dictionary app. Both the Dic Group and the Dic+MT Group also reported all the words they looked up in the dictionary app (although for data analysis we only focused on the target music words). It should be noted that the students used the MT apps for sentence translations only,

[1] The two apps are powered by neural machine translation technologies.

rather than using them like a dictionary. This was to ensure that overlapping app functions would not be a confounding factor to reduce the experimental validity.

Upon finishing the translation task, the instructor showed the correct translations and guided the students to read the sentences and translations three times. After that, the students translated the same 20 sentences again, as the immediate posttest, without access to any tool. At the end of the posttest, the students took a two-part survey. The first part prompted them to report their perceptions in four aspects (adapted from Dai & Wu, 2023). They rated on a 6-point scale (6 = very much; 1 = not at all) about the extent to which they believed that (a) the app was useful; (b) the app was easy to use; (c) they liked using the app; and (d) they intended to continue using the app for learning music words. The mean score of these four items was taken as the perception rating for a particular app. The second part of the questionnaire was voluntary, soliciting the students' qualitative comments on the perceived advantages and disadvantages of using the apps for learning music words.

In the third week, the students took an unexpected delayed posttest. They translated Set B sentences (see Section 3.2) without the help of any tool.

3.5 Variables and data analysis

The target music words in the tests were rated on the following scale: correct translation (1 point); partially correct translation (0.5 point); wrong or no translation (0 point). The translations were first graded by a research assistant and then checked by the first author. Inconsistency was resolved through the discussion between the first and the second authors until complete agreement was reached. As some data were not normally distributed, we conducted Kruskal-Wallis H tests to find out whether inter-group differences existed, followed up by Dunn's multiple comparisons as post-hoc tests (*p* values adjusted with the Holm's method). These comparisons showed whether the use of mobile dictionary and/or MT apps led to different retention rates of the (semi-)technical words among the three groups, thereby addressing RQ1.

To answer RQ2 about how the Dic Group and the Dic+MT Group differed in their dictionary consultation behaviors, we compared the numbers of target (semi-)technical words looked up by the two groups, using the Mann-Whitney test.

To examine student perceptions of the apps (RQ3), we conducted Mann-Whitney tests for the between-group comparisons of (a) dictionary app ratings between the Dic Group and the Dic+MT Group and (b) MT app ratings between the Dic+MT Group and the MT Group. To analyze the students' comments, we followed Miles et al.'s (2014) two-cycle coding protocol, first using the students' own words as *in vivo* codes and then synthesizing them into coherent thematic codes. Coding differences were resolved through discussion between the authors until consensus was reached. These thematic insights enabled us to obtain a richer understanding of student perceptions, complementing the quantitative survey results.

4. Results

4.1 Test results

Table 1 reports the medians (Mdn) and interquartile ranges (IQR) of the test results. Kruskal-Wallis H tests showed that the three groups differed significantly in the posttests for both semi-

technical words ($H(2) = 31.08$, $p < 0.001$) and technical words ($H(2) = 35.86$, $p < 0.001$). Dunn's multiple comparisons revealed that the Dic Group outperformed the Dic+MT Group, who in turn outperformed the MT Group, for both semi-technical and technical words (see Table 2 for detail).

The three groups also differed significantly in the delayed posttests for both semi-technical words ($H(2) = 22.48$, $p < 0.001$) and technical words ($H(2) = 29.07$, $p < 0.001$). Dunn's multiple comparisons indicated that the Dic Group outperformed the Dic+MT Group, who in turn outperformed the MT Group, for both semi-technical and technical words (see Table 3 for detail).

Overall, the posttest and the delayed posttest suggested that the retention rates differed significantly among the groups. As judged by the delayed posttest, the Dic Group had the best retention rates, with a median of 4 semi-technical words and 4 technical words, or 40% of the target music words. The Dic+MT Group was in-between, with a median of 3 semi-technical words and 2 technical words, or about 25% of the target words. The MT Group had the worst retention rates, with a median of 1 semi-technical word and 1 technical word, or 10% of the target words.

Table 1. Descriptive statistics for the posttest and delayed posttest

Groups	Posttest				Delayed posttest			
	Semi-technical words		Technical words		Semi-technical words		Technical words	
	Mdn	IQR	Mdn	IQR	Mdn	IQR	Mdn	IQR
Dic	8	3	8	2	4	3	4	4
Dic+MT	6	5	6	2.5	3	2.5	2	3
MT	3	2.5	4	2.5	1	1	1	2

Table 2. Dunn's multiple comparisons for the posttest

Comparisons	Semi-technical words		Technical words	
	Z	*p* (Holm-adjusted)	Z	*p* (Holm-adjusted)
Dic vs. Dic+MT	2.64	0.008	3.13	0.003
Dic+MT vs. MT	2.94	0.007	2.85	0.004
Dic vs. MT	5.57	< 0.001	5.99	< 0.001

Table 3. Dunn's multiple comparisons for the delayed posttest

Comparisons	Semi-technical words		Technical words	
	Z	*p* (Holm-adjusted)	Z	*p* (Holm-adjusted)
Dic vs. Dic+Trans	2.57	0.02	2.9	0.007
Dic+Trans vs. Trans	2.16	0.03	2.49	0.013
Dic vs. Trans	4.74	<0.001	5.39	<0.001

4.2 Look-up behaviors

Table 4 reports the descriptive statistics for the numbers of target words looked up in the dictionary app during the translation task. The Mann-Whitney tests showed that the Dic Group

looked up significantly more semi-technical words than the Dic+MT Group ($z = 4.03$, $p < 0.0001$). This was also true for the technical words ($z = 4.92$, $p < 0.0001$). The lookup rates of the Dic Group were twice as many as those of the Dic+MT Group. This might be because the Dic Group must translate the sentences from scratch. When they encountered new words or known words with new meanings, they must consult the dictionary app to facilitate their comprehension, without which the translation task would not have been possible. By contrast, the Dic+MT Group worked on the MT output. When the machine-translated texts read natural to them and/or they could not tell whether the translations were incorrect, they might skip consulting the dictionary app.

Table 4. Numbers of the target words looked up in the dictionary app

Group	Semi-technical words		Technical words	
	Mdn	IQR	Mdn	IQR
Dic	7	3	10	0
Dic+MT	4	4	5	5.5

4.3 Student perceptions of the apps

Table 5 reports the medians and IQRs of the perception ratings of the apps. Mann-Whitney tests showed that the Dic Group and the Dic+MT Group did not differ significantly ($z = 0.74$, $p = 0.46$) in their perceptions of the dictionary app. Similarly, the Dic+MT Group and the MT Group did not differ significantly ($z = 1.26$, $p = 0.21$) in their perceptions of the MT apps. These seem to show that the students were largely satisfied with the digital tools and the perception ratings did not differ. However, closer examination of their qualitative comments revealed some interesting patterns and nuanced differences. As the comments were voluntary, not all the students responded to the open-ended questions in the survey. Table 6 and Table 7 summarize the numbers of students who shared their perceived advantages and disadvantages of the apps.

Table 5. Perception ratings of the dictionary and MT apps

Group	Dictionary app		Translation app	
	Mdn	IQR	Mdn	IQR
Dic	5	1	*(not applicable)*	*(not applicable)*
Dic+MT	5	2	5	2
MT	*(not applicable)*	*(not applicable)*	4.75	0.75

Overall, "domain-specificity" was the most frequently reported advantage of the dictionary app by both the Dic Group and the Dic+MT Group (Table 6), although the percentage in the former was much higher than that in the latter (74.1% vs. 47.6%). Relatedly, "translation insensitive to domain-specificity" was the second most frequently cited disadvantage of the MT apps by both the Dic+MT Group and the MT Group (Table 7). These two groups also recognized "quick access" as the most appreciated feature and "inaccurate translation" as the most disfavored feature of the MT apps (both accounting for more than half of the responses in each group). These top-ranking commonalities might explain why the groups did not differ significantly in

their perception ratings, as reported in the previous paragraph. However, some differences (albeit low in frequency counts) were identified in their perceptions of the content, functionality, usability, and popularity of the apps. By comparing the comments across the three groups, we found that some students' perceptions of the dictionary app seemed to be influenced by the MT apps. Specifically, four students from the Dic+MT Group considered "accurate translation" as an advantage of the dictionary app (Table 6), which was not commented by the Dic Group at all. As such, it was possible that the MT apps sensitized some students from the Dic+MT Group to the accuracy of the dictionary app. In addition, three students from the Dic+MT Group saw "popularity" as an advantage of the MT apps (Table 7), while two cited "not popular" as a disadvantage of the dictionary app (Table 6). These comments were exclusive to the Dic+MT Group, again suggesting that some students might advertently pit one type of app against another in their perceptions.

Table 6. Voluntary comments about the dictionary app

Perceptions	Dic Group		Dic+MT Group	
Advantages	(n = 28)		(n = 21)	
Domain-specificity	20	71.4%	10	47.6%
Accurate translation	0	0%	4	19%
Rich lexicographical content	3	10.7%	1	4.8%
Search history	3	10.7%	0	0%
Quick access	0	0%	2	9.5%
User-friendly interface	1	3.6%	3	14.3%
Other	1	3.6%	1	4.8%
Disadvantages	(n = 5)		(n = 6)	
Layout	3	60%	3	50%
User-unfriendly interface	2	40%	1	16.7%
Not popular	0	0%	2	33.3%

Table 7. Voluntary comments about the MT apps

Perceptions	Dic+MT Group		MT Group	
Advantages	(n = 24)		(n = 24)	
Quick access	13	54.2%	18	75.0%
Popularity	3	12.5%	0	0%
Accurate translation	2	8.3%	1	4.2%
User-friendly interface	2	8.3%	2	8.3%
Enabled comprehension	2	8.3%	3	12.5%
Other	2	8.3%	0	0%
Disadvantages	(n = 21)		(n = 25)	
Inaccurate translation	14	66.7%	16	64%
Translation insensitive to domain-specificity	6	28.6%	6	24%
Other	1	4.8%	3	12%

5. Discussion

The test results showed that the three groups significantly differed in their retention of (semi-)technical music words. The Dic Group had the best retention rates, followed by the Dic+MT Group. The MT Group had the worst retention rates.

The Dic+MT Group was outperformed by the Dic Group for two possible reasons. First, the Dic+MT Group might not notice the problematic translations in the MT output. Thus, they did not check the meanings in the dictionary app and missed the learning opportunities, evidenced by their halved lookup rates compared to the Dic Group (see Table 4). As Webb and Nation (2017) aptly pointed out, "Although noticing by itself is not one of the deeper levels of quality of attention, it is still very important, because where we give our attention largely determines what will be learned" (p. 68). Framed from a CLT perspective, the Dic+MT Group, without noticing the deviated meanings in the MT output, did not devote their attentional resources to vocabulary learning. Further, when they indeed noticed the semantic issues and looked up the words in the dictionary app, they might be exposed to extraneous cognitive load not beneficial to learning. As indicated by the qualitative comments, only the students from the Dic+MT group singled out "accuracy" as a strength of the dictionary app and "popularity" as an acknowledged feature of the MT apps. These comments suggest that some students might expend mental resources on making comparisons between the MT and dictionary apps, thus possibly draining their germane cognitive load. In other words, exposure to two types of apps might not serve as "cognitive extension" but rather induce "cognitive distraction" for some students (Yoon, 2016, p. 222).

Of the three groups, the MT Group had the worst retention rates of the (semi-)technical words, suggesting that MT had a minimal effect on vocabulary retention, consistent with Klimova et al.'s (2023) observation. However, our findings are different from the positive, MT-enabled learning effects reported by Lo (2023, 2024). The discrepancy may be due to the research designs. The target words in Lo (2023, 2024) had unambiguous, correct translations in Google Translate. As such, the students in her studies did not need to identify or correct the MT errors. However, in the current study, the MT versions of the target words might be erroneous and the students might fail to notice the MT errors, thus not acting on the learning opportunities. Even when the students from the MT Group were aware of the problematic translations, given the lack of access to other materials (e.g., dictionaries), they could only infer the meanings of the music words from the co-texts, without being able to verify their hypotheses. This might be as far as their cognitive engagement could go and their mental resources could not be optimally devoted to the input germane to vocabulary learning. Conversely, since the MT output of the target words in Lo (2023, 2024) was always correct, the participants could optimally expend their germane cognitive load on learning the words.

Compared with the other two groups, the Dic Group had distinct advantages in terms of attentional resources and germane cognitive load. In the sentence translation task, the Dic Group worked from scratch and was very likely to notice the gap between their vocabulary knowledge and the word meaning most suitable for the sentential context. This prompted them to attend to the gaps by consulting the dictionary app as their only resort. Since they were not distracted by the MT output or unnecessary searches, they could maximize their germane

cognitive load and focus on the meanings of the music words. This could explain why they had the best retention rates among the three groups.

Regarding the perceptions, the students held a generally positive attitude towards the dictionary and MT apps, consistent with the findings in O'Neill's (2019a) and Almusharraf and Bailey (2023). The qualitative comments further elucidated why the students were receptive to these apps. Both the Dic Group and the Dic+MT Group recognized "domain-specificity" as the most appreciated feature of the dictionary app. Both the Dic+MT Group and the MT Group reported "quick access" as the most valued advantage of the MT apps. Interestingly, the students also pointed out "inaccurate translation" as the primary concern of the MT apps. However, the students' comments gravitated towards these top-ranking (dis)advantages, with much fewer comments on other app features (e.g., search history and user interface). These show that most students recognized a limited set of affordances and challenges inherent in the apps. In addition, we were somewhat surprised that no inter-group differences were found in the students' perception ratings, despite their significant inter-group differences in the retention rates of the target words. This performance-perception discrepancy suggests that the students had not developed a sophisticated understanding of the affordances of mobile apps and the ability to mindfully relate the affordances to their learning performances (detailed in the next section).

6. Implications

Our study has three pedagogical implications for the mobile-assisted learning of music words, and by extension, specialized words in other fields. First, our findings disfavor the use of MT by low-proficiency students for learning specialized vocabulary. We are not discounting the usefulness of MT for other purposes (such as facilitating English writing, Lee, 2020). However, MT may not be an ideal tool for low-proficiency students to learn (semi-)technical words. Our survey found that the students valued the quick access to information afforded by MT and held a generally positive attitude towards the MT apps, even though (a) they cited "inaccurate translation" and "translation insensitive to domain-specificity" as two glaring weaknesses and (b) the use of MT did not lead to the best retention rates. These show that (a) the students valued immediacy over quality when using MT (O'Neill, 2019a) and (b) student perceptions and performances were misaligned. These issues shall be addressed by ESP instructors so that students will develop a more critical attitude towards MT apps. Webb and Nation (2017) caution that exposure to incorrect word forms will interfere with the learning of forms. Similarly, based on our research findings, we maintain that exposure to erroneous meanings generated by MT possibly impedes the learning of specialized vocabulary, at least for low-proficiency students. This is because they may overestimate their ability to identify and address MT errors (Loock & Léchauguette, 2021) and may not crosscheck the MT output (Hellmich & Vinall, 2023). Even when they do follow up on the MT errors and consult the dictionary app, they may be exposed to increased extraneous cognitive load induced by making comparisons between the two apps. Our findings show that the learning effect is not ideal when germane cognitive load is not sufficiently devoted to the specialized words themselves.

Second, the students solely using the dictionary app had the best retention rates of both semi-technical and technical words, suggesting that the learning effect of dictionary apps may

be superior to that of MT apps (see also Almusharraf & Bailey, 2023). In addition, the students recognized the domain-specific content as the most valued feature afforded by the dictionary app. It should be noted that we only included the specialized meanings of the semi-technical words to reduce the students' extraneous cognitive load potentially caused by unnecessary searches. Peters and Fernández (2013) observe that semi-technical words are seldom included in specialized dictionaries. Thus, it is important for future lexicographical endeavors to cover the specialized meanings of semi-technical words in ESP dictionaries to maximize learning opportunities. For pedagogical endeavors, instructors need to train students to make the best use of the adaptability of dictionary apps. While it is difficult to customize the content in paper dictionaries, it is plausible to do so in some dictionary apps (such as the one used in this study). For instance, in the context of ESP learning, students can import ESP lexicographical databases and configure the content display so that specialized meanings are presented before (if not in the place of) the EGP meanings in the dictionary app. This would lower "the search-related information costs…(i.e. effort) related to the lookup acts" (Nielsen, 2008, p. 173) and subsequently free up cognitive resources germane to specialized vocabulary learning.

Third, "word consciousness training and activities" should be conducted (Nation, 2022, p. 108) so that students are sensitized to the deceptive transparency of semi-technical words. In our study, both the Dic Group and the Dic+MT Group were more likely to look up the technical words than the semi-technical words. This suggests that they might mistake the unspecialized meaning for the specialized meaning. As students may stick to meanings familiar to them (Peters, 2020), ESP instructors need to heighten student awareness of the polysemous and opaque nature of semi-technical words (Watson Todd, 2017). For instance, following a data-driven learning approach (Skoufaki & Petrić, 2021), instructors can select two sample sentences of a semi-technical word: one from an EGP corpus exemplifying the unspecialized meaning, the other from an ESP corpus illustrating the specialized meaning. Students can be guided to examine how the meanings are differently primed by contextual cues. Then, they can apply this knowledge to more pairs of sample sentences to develop their sensitivity to semi-technical words.

7. Conclusion

Our study contributes to MALL research by examining how dictionary and MT apps differ in their effects on the retention of (semi-)technical music words. It was found that the Dic Group was the best, followed by the Dic+MT Group. The MT Group was the poorest. Additionally, the Dic Group looked up twice as many target words as the Dic+MT Group. From the CLT perspective, these results suggest that the Dic Group had the greatest germane cognitive load because (a) when translating the sentences from scratch, they were more likely to notice the unknown/unfamiliar words and consult the dictionary app to figure out the meanings; and (b) the look-up acts could incur deliberate attention to word features and engage them in "thoughtful processing of" lexical information germane to vocabulary learning (Webb & Nation, 2017, p. 120). Notwithstanding the inter-group differences in vocabulary retention, the perception ratings did not differ between the groups, showing that student perceptions and performances were disconnected. Thus, ESP instructors should train students to be critically reflexive users of mobile apps in learning specialized words and engage them in word

consciousness activities to develop their sensitivity to the opaque nature of semi-technical words.

Despite its meaningful insights into mobile-assisted vocabulary learning, the current study is not without limitations. One potential limitation is that the study spanned a relatively short period of time, although it should be noted that short treatment durations were typical in similar previous studies (see Lin & Lin, 2019) and a meta-analysis has shown that "technology enhanced L2 vocabulary learning regardless of how long the study intervention lasted" (Hao et al., 2021, p. 662). To obtain a fuller picture, future studies can adopt a longitudinal design and examine whether the differential learning effects among the mobile-assisted conditions will become more or less pronounced over time. Another limitation of the study is its sole focus on low-proficiency students. It will be interesting for future studies to find out whether high-proficiency students will be more skillful in addressing MT errors (Lee, 2023) and whether crosschecking with dictionary apps will increase their extraneous cognitive load or not. Directly comparing how different proficiency groups use mobile apps will reveal preferences and needs that are common and/or unique to their proficiency levels. These insights will enable us to offer incisive instruction to develop students' ability to perceive and act on affordances of mobile apps to advance their specialized vocabulary learning.

Appendices

Appendix A. Target music words

Semi-technical words	**Technical words**
arrangement	vibrato
bar	harmonium
flat	tambourine
march	ostinato
movement	baritone
register	cymbal
staff	falsetto
number	oratorio
bridge	carillon
round	plectrum

Appendix B. Chi-squared test of the number of students using translation apps

Group	Youdao	Baidu	Chi-squared test
Dic+MT Group	15	16	$\chi^2 = 0.258$ ($p = 0.611$)
MT Group	17	14	

References

Almusharraf, A., & Bailey, D. (2023). Machine translation in language acquisition: A study on EFL students' perceptions and practices in Saudi Arabia and South Korea. *Journal of Computer Assisted Learning*, *39*(6), 1988-2003.

Barcroft, J. (2009). Effects of synonym generation on incidental and intentional L2 vocabulary learning during reading. *TESOL Quarterly*, *43*(1), 79-103.

Barenberg, J., Berse, T., Reimann, L., & Dutke, S. (2021). Testing and transfer: Retrieval practice effects across test formats in English vocabulary learning in school. *Applied Cognitive Psychology*, *35*(3), 700-710.

Cancino, M., & Panes, J. (2021). The impact of Google Translate on L2 writing quality measures: Evidence from Chilean EFL high school learners. *System*, *98*, 102464.

Cervetti, G. N., Hiebert, E. H., Pearson, P. D., & McClung, N. A. (2015). Factors that influence the difficulty of science words. *Journal of Literacy Research*, *47*(2), 153-185.

Chon, Y. V., Shin, D., & Kim, G. E. (2021). Comparing L2 learners' writing against parallel machine-translated texts: Raters' assessment, linguistic complexity and errors. *System*, *96*, 102408.

Chung, E. S., & Ahn, S. (2022). The effect of using machine translation on linguistic features in L2 writing across proficiency levels and text genres. *Computer Assisted Language Learning*, *35*(9), 2239-2264.

Dai, Y., & Wu, Z. (2023). Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: A mixed-methods study. *Computer Assisted Language Learning*, *36*(5-6), 861-884.

Dai, Y., Wu, Z., & Xu, H. (2019). The effect of types of dictionary presentation on the retention of metaphorical collocations: Involvement load hypothesis vs. cognitive load theory. *International Journal of Lexicography*, *32*(4), 411-431.

Dang, T. N. Y., Lu, C., & Webb, S. (2022). Incidental learning of single words and collocations through viewing an academic lecture. *Studies in Second Language Acquisition*, *44*(3), 708-736.

Deignan, A., & Love, R. (2021). Using corpus methods to identify subject specific uses of polysemous words in English secondary school science materials. *Corpora*, *16*(2), 165-189.

Dziemianko, A. (2018). Electronic dictionaries. In P. A. Fuertes-Olivera (Ed.), *The Routledge handbook of lexicography* (pp. 663-683). Routledge.

Dziemianko, A. (2022). The usefulness of graphic illustrations in online dictionaries. *ReCALL*, *34*(2), 218-234.

Fredholm, K. (2019). Effects of Google translate on lexical diversity: Vocabulary development among learners of Spanish as a foreign language. *Revista Nebrija de Lingüística Aplicada a la Enseñanza de las Lenguas*, *13*(26), 98-117.

Gablasova, D. (2015). Learning technical words through L1 and L2: Completeness and accuracy of word meanings. *English for Specific Purposes*, *39*, 62-74.

Garcia, I., & Pena, M. I. (2011). Machine translation-assisted language learning: writing for beginners. *Computer Assisted Language Learning*, *24*(5), 471-487.

Hao, T., Wang, Z., & Ardasheva, Y. (2021). Technology-assisted vocabulary learning for EFL learners: A meta-analysis. *Journal of Research on Educational Effectiveness*, *14*(3), 645-667.

Haynes, M., & Baker, I. (1993). American and Chinese readers learning from lexical familiarization in English texts. In T. Huckin, M. Haynes, & J. Coady (eds.), *Second language reading and vocabulary learning* (pp. 130-152). Ablex.

Hellmich, E. A., & Vinall, K. (2023). Student use and instructor beliefs: Machine translation in language education. *Language Learning & Technology*, *27*(1), 1-27.

Klimova, B., Pikhart, M., Benites, A. D., Lehr, C., & Sanchez-Stockhammer, C. (2023). Neural machine translation in foreign language teaching and learning: a systematic review. *Education and Information Technologies*, *28*(1), 663-682.

Lee, S. M. (2020). The impact of using machine translation on EFL students' writing. *Computer Assisted Language Learning*, *33*(3), 157-175.

Lee, S. M. (2023). The effectiveness of machine translation in foreign language education: a systematic review and meta-analysis. *Computer Assisted Language Learning*, *36*(1-2), 103-125.

Levy, M., & Steel, C. (2015). Language learner perspectives on the functionality and use of electronic language dictionaries. *ReCALL*, *27*(2), 177-196.

Lin, J. J., & Lin, H. (2019). Mobile-assisted ESL/EFL vocabulary learning: A systematic review and meta-analysis. *Computer Assisted Language Learning*, *32*(8), 878-919.

Liu, T. C., Fan, M. H. M., & Paas, F. (2014). Effects of digital dictionary format on incidental acquisition of spelling knowledge and cognitive load during second language learning: Click-on vs. key-in dictionaries. *Computers & Education*, *70*, 9-20.

Lo, S. (2023). Neural machine translation in EFL classrooms: learners' vocabulary improvement, immediate vocabulary retention and delayed vocabulary retention. *Computer Assisted Language Learning*, 1-20.

Lo, S. (2024). The effects of NMT as a de facto dictionary on vocabulary learning: a comparison of three look-up conditions. *Computer Assisted Language Learning*, 1-21.

Loock, R., & Léchauguette, S. (2021). Machine translation literacy and undergraduate students in applied languages: report on an exploratory study. *Revista Tradumàtica*, (19), 204-225.

Lu, H., Zhang, Y., & Hao, X. (2020). The contribution of cognitive linguistics to the acquisition of polysemy: A dictionary entry-based study with Chinese learners of English. *International Journal of Lexicography*, *33*(3), 306-336.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd edition). Sage.

Nielsen, S. (2008). The effect of lexicographical information costs on dictionary making and use. *Lexikos*, *18*(1), 170-189.

Nation, I. S. P. (2022). *Learning vocabulary in another language* (3rd edition). Cambridge University Press.

Niño, A. (2008). Evaluating the use of machine translation post-editing in the foreign language class. *Computer Assisted Language Learning*, *21*(1), 29-49.

Niño, A. (2009). Machine translation in foreign language learning: Language learners' and tutors' perceptions of its advantages and disadvantages. *ReCALL*, *21*(2), 241-258.

O'Neill, E. M. (2019a). Online translator, dictionary, and search engine use among L2 students. *CALL-EJ: Computer-Assisted Language Learning–Electronic Journal, 20*(1), 154-177.

O'Neill, E. M. (2019b). Training students to use online translators and dictionaries: The impact on second language writing scores. *International Journal of Research Studies in Language Learning, 8*(2), 47-65.

Peters, E. (2020). Factors affecting the learning of single-word items. In S. Webb (ed.). *The Routledge handbook of vocabulary studies* (pp. 125-142). Routledge.

Peters, P., & Fernández, T. (2013). The lexical needs of ESP students in a professional field. *English for Specific Purposes, 32*(4), 236-247.

Skoufaki, S., & Petrić, B. (2021). Exploring polysemy in the Academic Vocabulary List: A lexicographic approach. *Journal of English for Academic Purposes, 54*, 101038.

Slatyer, H., & Forget, S. (2020). Digital translation: its potential and limitations for informal language learning. In M. Dressman & R.W. Sadler (Eds.). *The handbook of informal language learning* (pp. 439-456). John Wiley & Sons Ltd.

Smith, A., & Davis, A. (2018). Disambiguating the use of common terms across related medical fields: the problem of intervention. *Lexicography, 4*(1), 63-80.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22*(2), 123-138.

Sweller, J., Kalyuga, S. & Ayres, P. (2011). *Cognitive load theory*. Springer.

Tarp, S. (2010). Functions of specialised learners' dictionaries. In P.A. Fuertes-Olivera (Ed.). *Specialised dictionaries for learners* (pp. 39-53). de Gruyter.

Watson Todd, R. (2017). An opaque engineering word list: Which words should a teacher focus on?. *English for Specific Purposes, 45*, 31-39.

Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.

Yoon, C. (2016). Concordancers and dictionaries as problem-solving tools for ESL academic writing. *Language Learning & Technology, 20*(1), 209-229.

Yüksel, H. G., Mercanoğlu, H. G., & Yılmaz, M. B. (2022). Digital flashcards vs. wordlists for learning technical vocabulary. *Computer Assisted Language Learning, 35*(8), 2001-2017.