An integrated corpus-based text mining approach used to process military technical information for facilitating EFL troopers’ linguistic comprehension: US anti-tank missile systems field manual as an example

LC Chen¹,2, KH Chang³,⁴* and SC Yang²
¹ Department of Foreign Languages, R.O.C. Military Academy, Kaohsiung 830, Taiwan.
² Institute of Education, National Sun Yat-Sen University, Kaohsiung 804, Taiwan.
³ Department of Management Sciences, R.O.C. Military Academy, Kaohsiung 830, Taiwan.
⁴ Institute of Innovation and Circular Economy, Asia University, Taichung 413, Taiwan.

Submitted: 02 September 2020; Revised: 28 December 2020; Accepted: 26 March 2021

Abstract: Military knowledge is an uncommon research field and is often classified as confidential information. Furthermore, when US military knowledge is adopted by English as a foreign language (EFL) countries, properly interpreting military texts brings about challenges. Taking Asian militaries as examples of EFL countries, not every trooper has sufficient English proficiency and capability to read and comprehend complicated military knowledge databases. In addition, under limited training time and lack of suitable reference materials, it is difficult to popularise and improve the efficiency of the courses that study US field manuals (FMs), which are important books that introduce US military combat tactics and strategies, military operation procedures, weapon systems, and others. Nevertheless, in many EFL countries, English learning is integrated into the education system to promote internationalisation and enhance global competitiveness. Thus, the English proficiency of nationals in most EFL countries is not negligible. Based on these considerations, this paper discusses the integration of the corpus software and cooperation of linguists and military experts to conduct syntax analysis and taxonomy of military terminology to enable EFL troopers with non-excellent English proficiency to understand the intricate US military domain knowledge and develop the military corpus as an auxiliary language training material. The US Army FMs of anti-tank missile systems are adopted as an empirical example to illustrate the proposed approach. Analytical findings will become critical reference indicators for defence language institutes (DLI) of EFL militaries in developing military English training materials and for processing military information.

Keywords: Anti-tank missile systems, Asian militaries, corpus software, military corpus, military terminology, syntax analysis.

INTRODUCTION

Corpus-based analytical approaches are considered as big data analysis; its sources of big data (Christ et al., 2019) are natural languages (NLs) that are compiled to become corpus from enormous texts and discourses (Koops & Lohmann, 2015; Brindle, 2016; Beller & Bender, 2017; Chen et al., 2020). Statistics-based algorithm corpus analysis studies are indispensable in today’s digital era. Due to the developments in computer technology, large amounts of texts which include articles, news reports and discourses, among others can be stored electronically in computers (Ferguson, 2001; Daskalovska, 2015; Coats, 2019). Thus, corpus programs directly process text information, via computers for further analysis (Chen & Chang, 2019; Smalheiser et al., 2019). In terms of military information, it has sophisticated domain knowledge and scientific techniques, moreover, it is difficult to collect and is often categorised as confidential data (Trembach, 2019). Hence, this kind of analysis is pertinent to countries where English is a foreign language (EFL) and weaponries and military operation techniques are mainly adopted from the US military. A chunk of the military domain knowledge might appear ambiguous to troopers who use English as a foreign language (EFL) troopers. Seeking proper information processing toolkits is necessary for improving the efficiency of EFL military information processing. To conquer the problems resulting from language difficulties, computer-assisted language learning (CALL)
has been identified as a key tool for improving foreign languages acquisition since the invention of computers.

CALL is an educational research term that discusses the interrelationship between language teaching, learning, and digital technologies. The core of CALL is to adopt various computer-based technological resources to enhance the efficacy of language teaching and learning. With modern technological advancement, CALL applications have also evolved from single unit systems to multi-function systems that are more complex and much closer to simulating real-life situations (Gonen, 2019). For example, Hsieh et al. (2017) embedded a social media app, LINE, into a foreign language course. The results of their study showed that a social communication program would be an affordable, ubiquitous, and easy to use pedagogical assistant tool for stimulating students’ learning motivation and improving students’ language learning efficacy. Harvey-Scholes (2018) used the computer program, N-gram method, to detect native Spanish speakers’ English writing errors for improving EFL students’ writing proficiency. Cheng and Tsai (2019) assimilated head-mounted displays used for creating virtual realities into field trip education. They found that this pedagogical method that embedded novel technology can effectively enhance students’ learning motivation and reduce exam anxiety. CALL applications have been gradually utilised by language teachers (or instructors) as digitalised coworkers in their language classrooms. Nowadays, CALL is more learner-centered, with the availability of learning materials for providing autonomous and continuous learning approaches (Lai et al., 2016). The aforementioned studies mainly discuss the CALL approaches used in language courses for enhancing EFL learners’ learning achievements. When languages are used for specific purposes, the problems to be solved may be slightly different. US military information consists of linguistic and domain knowledge, both English and military; thus, it is considered as an example of English for Specific Purposes (ESP) case.

ESP mainly discusses English linguistic knowledge that is applied in specific domains and pedagogical approaches or self-learning methods for making EFL learners acquire disciplinary literacy. Namely, ESP curricula and learning tools’ developments lean towards the domains’ needs. Disciplinary literacy, as Zygouris-Coe (2012) defined it, is mainly centered on the learning domain knowledge and putting the domain knowledge into an actual environment, where the language use is English. It emphasises a middle-ground between English acquisition and its application to domain knowledge. Moreover, in order to enable ESP learners to use English for gaining expertise in domains, concept-embedding words and lexical bundles (LBs) inevitably need to be extracted and researched. Shanahan and Shanahan (2017), proposed a pedagogical strategy for learning ESP. They proposed that learners should focus on vocabulary for general purposes at the initial stage. When the learners attain certain comprehensive levels of vocabulary and grammar, they will begin to acquire the requisite vocabulary to handle schools’ academic needs. Finally, the learners would acquire the specific linguistic knowledge for a discipline and expertise in a domain. Based on the theories postulated by Zygouris-Coe (2012) and Shanahan and Shanahan (2017), it is obvious that ‘vocabularies' are the essential elements in ESP research. Furthermore, retrieving concept-embedding words and LBs such as terminology, technical words, and domain-specific phrases are also essential tasks. In ESP research cases, many pedagogical approaches and knowledge processing approaches are applied in surmounting language barriers that are particularly caused by EFL (Flowerdew, 2000; Zygouris-Coe, 2012). However, there are no absolute advantages for English native speakers in gaining domain knowledge (Zygouris-Coe, 2012; Shanahan & Shanahan, 2017; Viswanathan et al., 2020). For instance, Derbentseva et al. (2007) structured concept maps to illustrate intricate domain knowledge for improving students’ subject learning proficiency in Canada, where English is the first language. In addition, in EFL environments, corpus-based approaches are popular because its results align more with domain experts and linguists' expectations. Li et al. (2019) reviewed the literature that used corpus-based methods to analyse tourism information and pointed out the intimate correlation between corpus analysis and big data analysis. Li (2016) used Wordsmith tool, the most popular corpus program, to extract the word list and keyword list from the corpus of JRC-Acquis (EN) in order to identify vague terms that are used in legal documents. Munoz (2015), in addition to use the corpus program to retrieve the keyword list from the agricultural corpus, also conducted a taxonomy of keywords so that the data output will obtain greater benefits in ESP courses’ development and ESP learning. Other research activities that employed corpus-based approaches to probe linguistic patterns in texts of different professional disciplines such as medicine (Li et al. 2019; Siefridt et al., 2020), engineering (Liu & Han, 2015; Nekrasova-Beker, 2019), linguistic and language education (Henry & Roseberry, 2001; Green, 2019; Kim & Nam, 2019), and others, have also significantly resolved complex ESP cases.

Recently, statistics-based algorithm corpus programs have been advanced gradually by modern computer technologies, but the limitations of corpus-
METHODOLOGY

For military domain, military knowledge embraces professional and uncommon genre types, terminologies and scientific knowledge that cause civilians or even military personnel some difficulty in receiving information. Thus, the proposed approach integrates machine processing and military experts’ annotation to process military corpus for retrieving domain knowledge, inducting genre types, and conducting military TISL taxonomy.

The proposed approach can be divided into two phases, machine process, and manual annotation, and covers seven steps (Figure 1). Steps 1 to 3 belong to Phase I - machine process; AntConc 3.5.8 (Anthony, 2019) is the primary analytical corpus software to process the target corpus. Steps 4 to 6 belong to Phase II - manual annotation. In step 5, linguists and domain experts conduct syntax analysis by clustering a function word list. In step (6), linguists and domain experts conduct the military TISL taxonomy based on checking the word list and keyword list, checking LBs of tokens, and checking concordance lines of abbreviations and acronyms. Step 7 clusters the results in step 5 and 6 for military ESP training courses. Detailed descriptions and illustrations of each step is introduced as follows:

Step 1: Creating the corpus

Although Antconc, Wmatrix, and Sketch Engine have similar functions in corpus analysis, nevertheless, considering the budgets and the necessity of internet, we would prefer Antconc because of its competitive advantages, such as affordable (i.e. free to access), ubiquitous (i.e. can be used anywhere), and easy (i.e. concise operative interfaces). Moreover, it can be operated without installation or an internet connection. Hence, this paper adopts AntConc 3.5.8 (Anthony, 2019), as the primary corpus software to process the target corpus.

Once textual data are collected, data conversion (transforming data from .docx, or .pdf into .txt) is inevitable because AntConc 3.5.8 (Anthony, 2019) only accepts particular data formations (i.e. .txt, UTF-8).

Step 2: Word list creation

Once all input data is prepared for analysis, users will choose the ‘Word List’ section and click the ‘Start’ button to generate the word list of the input corpus, and record it for further analysis.
Step 3: Keyword list creation

Normally, the algorithm of keyword list generator based on a log likelihood test to compare two corpus data, input corpus data and reference corpus data. The software utilises statistical algorithm to find words with high frequency in the input corpus but with low frequency in the benchmark corpus to compute their keyness values, to identify keywords. Keywords are considered significant features of the input corpus.

The selection of the benchmark corpus needs to base on genre types, namely, those two corpora have to have different genres (i.e. specific purposes vs. general purposes). Hence, the biggest and the most adopted general purposes genre type corpora include the corpus of contemporary American English (COCA) and the British National Corpus (BNC). Those corpora provide free access, and are ideal benchmark corpus data (e.g., Li, 2016).

After the word list is created and the benchmark corpus, input users will select the ‘Keyword List’ section and click the ‘Start’ button on the corpus software to generate the keyword list of the input corpus and record it for further analysis.

Step 4: Gathering related experts

Experts with linguistic analysis and military expertise are gathered to conduct the following procedures. All results from the corpus software, such as word list, keyword list, lexical bundles (LBs) of tokens and keywords, and so on, need to be analysed by experts in this step. Linguistic experts are expected to interpret genres of texts, while domain experts are expected to retrieve domain knowledge.

Step 5: Syntax analysis

In this step, the gathered experts, based on the word list that created in step 2, cluster a high-frequency function word list. Function words may seem literally meaningless; nevertheless, those are critical elements for structuring sentences, paragraphs, and even articles. Thus, high-frequency function words are critical clues for implementing syntax analysis.

Step 6: Military TISL taxonomy

In order to make extracted military TISL more meaningful for EFL troopers, in this step, the gathered experts will check the word list and keyword list (results in step 2 & 3) to extract military TISL, check LBs of tokens to avoid missing critical phrase-style terminologies, and check concordance lines of abbreviations and acronyms to retrieve the complete LBs and hidden meanings of abbreviations and acronyms. Eventually, the gathered experts will re-categorise military TISL.

Step 7. Military pedagogical applications

Results in step 5 and step 6 will become important in ESP training materials for EFL troopers before they enter the actual weaponry training.

RESULTS AND DISCUSSION

Overview of the military corpus data

In this study, the compiled military corpus data includes FM 3-23.25 Light anti-Armor Weapons (LAW) (US Department of the Army Washington, DC., 2001), FM 3-22.34 TOW Weapon System (US Department of the Army Washington, DC., 2003), FM 3-22.37 JAVELIN - Close Combat Missile System, Medium (US Department of the Army Washington, DC., 2008). Those weapon systems have been widely adopted by many countries and their FMs embrace the most complete technical information of the US Army current anti-tank weapon systems.

In the aforementioned military technical FMs, terms that are used for detailed weaponry specification, operating procedures, and tactical usages are introduced. Even if EFL speakers who have high proficiency in English research FMs, the interpretation of results may still cause information distortion because they lack military background knowledge. To verify the proposed approach in analysing military texts, the researchers adopted the compiled military corpus as an empirical example for importing to the proposed two-phase approach. The corpus contained three technical books (i.e. FMs), 5,346 word types and 108,605 tokens.

| Missile types                  | FM serial number | Word types | Tokens | TTR   |
|-------------------------------|------------------|------------|--------|-------|
| Javelin Missile               | FM 3-22.37       | 3,294      | 46,838 | 7.03 %|
| TOW Missile                   | FM 3-22.34       | 3,430      | 38,541 | 8.90 %|
| Light Anti-armor Weapon (LAW) | FM 3-23.25       | 2,518      | 23,226 | 10.84 %|
| Whole corpus                  |                  | 5,346      | 108,605| 4.92 %|

Table 1: Lexical characteristics of the target corpus
Its type/token ratio (TTR) is 4.92% (Table 1). FMs were segmented into each chapter (as sub-corpus), as it allowed the researchers to easily identify concordance plots and the etymology of words in the manuals. FMs’ figures, references and tables were eliminated. The elements of analysis were tokens, clusters, and concordances. In addition, this paper chose COCA as the benchmark corpus. COCA is the largest (contains 9,412,521 words) and genre-equivalent corpus of contemporary American English. It contains diverse texts which include discourses, fictions, newspapers, magazines, academic papers, and so on. Thus, using COCA as the reference corpus would be an ideal way to retrieve keywords from the target corpus.

**Resulting data of machine processing in Phase I**

In Phase I, AntConc 3.5.8 (Anthony, 2019) analysed the target corpus and generated the word list and keyword list. The raw data results are described as follows.

(1) Generating word list

The word list is the data resulting from step 2. The corpus program uses its statistic-based algorithm to integrate and count tokens’ frequency and to rank tokens. The word list indicated 5,346 words which ranked in frequency from high to low (see Table 2). High frequency words can be considered as the core elements of the target corpus. Moreover, low frequency words can also be considered as unique features of the target corpus.

**Table 2: An illustrative example of generating word list (partial data)**

| Rank | Frequency | Tokens | Rank | Frequency | Tokens | Rank | Frequency | Tokens |
|------|-----------|--------|------|-----------|--------|------|-----------|--------|
| 1    | 10988     | the    | 35   | 400       | s      | 68   | 205       | indicator |
| 2    | 3267      | and    | 36   | 393       | by     | 69   | 204       | during    |
| 3    | 2667      | to     | 37   | 390       | range  | 70   | 202       | use       |
| 4    | 2348      | of     | 38   | 377       | if     | 71   | 195       | d         |
| 5    | 2249      | a      | 39   | 376       | e      | 72   | 191       | one       |
| 6    | 1530      | is     | 40   | 368       | targets| 73   | 188       | right     |
| 7    | 1298      | in     | 41   | 339       | vehicle| 74   | 183       | battery   |
| 8    | 1018      | on     | 42   | 330       | can    | 75   | 183       | time      |
| 9    | 1007      | figure | 43   | 328       | should | 76   | 175       | two       |
| 10   | 889       | for    | 44   | 317       | round  | 77   | 171       | left      |
| 11   | 865       | gunner | 45   | 312       | must   | 78   | 168       | vehicles  |
| 12   | 865       | target | 46   | 308       | b      | 79   | 167       | these     |
| 13   | 814       | or     | 47   | 291       | unit   | 80   | 166       | trigger   |
| 14   | 718       | be     | 48   | 285       | weapon | 81   | 164       | each      |
| 15   | 653       | are    | 49   | 284       | table  | 82   | 164       | rear      |
| 16   | 643       | with   | 50   | 272       | used   | 83   | 160       | forward   |
| 17   | 633       | at     | 51   | 257       | light  | 84   | 159       | enemy     |
| 18   | 615       | missile| 52   | 253       | launcher| 85   | 157       | squad     |
| 19   | 581       | from   | 53   | 253       | meters | 86   | 156       | engagement |
| 20   | 565       | it     | 54   | 250       | launch | 87   | 152       | its       |
| 21   | 544       | position| 55   | 243       | switch | 88   | 151       | IR        |
| 22   | 517       | training| 56   | 235       | sight  | 89   | 150       | antiarmor |
| 23   | 514       | TOW    | 57   | 230       | system | 90   | 150       | no        |
| 24   | 509       | when   | 58   | 226       | area   | 91   | 147       | up        |
| 25   | 504       | as     | 59   | 222       | has    | 92   | 145       | c         |
| 26   | 492       | an     | 60   | 219       | gunnery| 93   | 145       | cover     |
| 27   | 471       | CLU    | 61   | 219       | seeker | 94   | 143       | day       |
| 28   | 457       | m      | 62   | 219       | will   | 95   | 142       | weapons   |
| 29   | 456       | that   | 63   | 217       | section| 96   | 141       | end       |
| 30   | 451       | fire   | 64   | 214       | may    | 97   | 141       | NVS       |
| 31   | 430       | Javelin | 65   | 210       | he     | 98   | 141       | other     |
| 32   | 409       | firing | 66   | 210       | trainer| 99   | 139       | using     |
| 33   | 404       | not    | 67   | 208       | all    | 100  | 138       | place     |
| 34   | 404       | this   |
(2) Generating Keyword list

The keyword list is the data output from step 3. The mechanism of generating keyword list is that the corpus software calculates ‘keyness’ of words by its algorithm, likelihood test, to find words that frequently appear in the target corpus but infrequently appear in the benchmark corpus. The keyword list, in this case (log-likelihood test (4-term), p < 0.05 (+ Bonferroni), covered 1,185 words and showed more specific words of the target corpus. In addition, it allowed us to filter function words or more generally-use words (Table 3).

Table 3: An illustrative example of generating keyword list (partial data)

| Rank | Freq. | Keyness | Keywords | Rank | Freq. | Keyness | Keywords |
|------|-------|---------|----------|------|-------|---------|----------|
| 1    | 865   | 7646    | gunner   | 51   | 82    | 679.91  | MGS      |
| 2    | 863   | 5838.75 | target   | 52   | 312   | 642.31  | must     |
| 3    | 1007  | 5675.24 | figure   | 53   | 130   | 634.9   | procedures |
| 4    | 615   | 5024.43 | missile  | 54   | 80    | 633.3   | qualification |
| 5    | 10988 | 4802.45 | the      | 55   | 89    | 632.97  | infrared |
| 6    | 514   | 4289.22 | TOW      | 56   | 226   | 630.88  | area     |
| 7    | 471   | 4216.11 | CLU      | 57   | 70    | 626.34  | FTT      |
| 8    | 430   | 3761.36 | JAVELIN  | 58   | 70    | 626.34  | NFOV     |
| 9    | 409   | 3029.92 | firing   | 59   | 98    | 621.61  | armor    |
| 10   | 544   | 2694.15 | position | 60   | 128   | 617.97  | mode     |
| 11   | 517   | 2652.34 | training | 61   | 68    | 608.45  | WFOV     |
| 12   | 368   | 2519.37 | targets  | 62   | 70    | 595.31  | loader   |
| 13   | 253   | 2240.87 | launcher | 63   | 142   | 586.15  | weapons  |
| 14   | 451   | 1977.8  | fire     | 64   | 97    | 585.58  | consists |
| 15   | 219   | 1937.1  | gunnery  | 65   | 115   | 581.38  | combat   |
| 16   | 390   | 1914.39 | range    | 66   | 63    | 563.7   | TFTT     |
| 17   | 339   | 1913.54 | vehicle  | 67   | 78    | 558.86  | connector |
| 18   | 219   | 1897.59 | seeker   | 68   | 85    | 555.15  | brightness |
| 19   | 285   | 1794.99 | weapon   | 69   | 125   | 551.9   | display  |
| 20   | 253   | 1737.53 | meters   | 70   | 115   | 541.1   | engage   |
| 21   | 291   | 1582.35 | unit     | 71   | 109   | 536.13  | soldier  |
| 22   | 317   | 1520.29 | round    | 72   | 61    | 535.59  | TGT      |
| 23   | 250   | 1498.35 | launch   | 73   | 70    | 534.48  | warhead  |
| 24   | 205   | 1498.01 | indicator| 74   | 63    | 533.49  | misfire  |
| 25   | 210   | 1428.04 | trainer  | 75   | 59    | 527.91  | firer    |
| 26   | 243   | 1416.54 | switch   | 76   | 75    | 498.65  | tactical |
| 27   | 150   | 1342.27 | antiarmor| 77   | 97    | 498.05  | tasks    |
| 28   | 235   | 1287.18 | sight    | 78   | 76    | 493.15  | ammunition |
| 29   | 141   | 1261.72 | NVS      | 79   | 71    | 487.13  | stationary |
| 30   | 151   | 1228.28 | IR       | 80   | 136   | 486.88  | conditions |
| 31   | 133   | 1190.13 | FOV      | 81   | 109   | 484.89  | command  |
| 32   | 376   | 1108.08 | e        | 82   | 108   | 483.59  | positions |
| 33   | 183   | 1058.92 | battery  | 83   | 87    | 481.38  | adjust   |
| 34   | 166   | 1018.3  | trigger  | 84   | 91    | 474.08  | tables   |
| 35   | 157   | 980.12  | squad    | 85   | 99    | 466.5   | objects  |
| 36   | 308   | 962.79  | b        | 86   | 129   | 464.91  | leader   |
| 37   | 156   | 944.72  | engagement| 87   | 230   | 464.56  | system   |
| 38   | 105   | 939.55  | nightsight| 88   | 272   | 459.41  | used     |
| 39   | 217   | 899.51  | section  | 89   | 145   | 456.39  | cover    |
| 40   | 284   | 892.83  | table    | 90   | 58    | 455.1   | platoon  |

Continued in page 409
Resulting data of experts’ annotations in Phase II

(1) Gathering related experts

For many EFL countries’ military, US military FMs are highly complex and critical because they involve a foreign language and military domain knowledge. Even if the corpus program is able to categorise and to process the target corpus, the contribution of raw data results to EFL troopers remains low. Thus, the researchers gathered related experts including linguists, military experts, and experts in performance evaluations (see Table 4) and appointed an assessment team to operate the analytical program in Phase I and to optimise the data results in Phase II.

Table 4: Introduction of linguistic and domain experts

| Researchers | Academic position | Academic specialties | Military specialty | Military rank | Years in active duty |
|-------------|-------------------|----------------------|--------------------|---------------|---------------------|
| Expert 1    | Professor         | Fuzzy Logic, Soft Computing, Applied Mathematics, Military Operations Research | Artillery          | Colonel       | 23 years            |
| Expert 2    | Assistant Professor | Performance Evaluation, Military Operations Research | Artillery          | Lieutenant Colonel | 19 years           |
| Expert 3    | Lecturer          | Corpus Linguistics, Computer-assisted language learning (CALL), Natural Language Processing (NLP) | Infantry           | Major         | 9 years             |
| Expert 4    | Weaponry Instructor | Management Science, also specialised in LAW, Javelin missile, TOW missile systems | Infantry           | Major         | 9 years             |

(2) Syntax analysis

According to the word list from step 2, all tokens were ranked by its frequency (refer to example data on Table 2). High frequency words were considered as important indicators for identifying core linguistic patterns in the target corpus. The word list featured many high-frequency function words. Those words may seem unrelated to the domain and easily be neglected; nevertheless, those are important elements of grammatical structures such as auxiliary verbs, prepositions, conjunctions, pronouns, and so on. Grammatical structures may confuse EFL troopers in their attempt at understanding FMs. Thus, the experts retrieved function words from the range of the top 500 high frequency words in the corpus of US Army antitank weapons FMs. The words were categorised into eight groups based on their grammatical functions (Table 5), then outlined the following linguistic evidences to conduct syntax analysis, for giving EFL military personnel important linguistic insights before they involve research or training courses in US anti-tank missile systems.

Group 1. “To-infinitive” and “for” represent purposes and reasons:

When “to-infinitive” clauses are placed after nouns or noun phrases, they indicate what the things refer to or intend to do. Using the target corpus as examples, the researchers discovered that “to-infinitive” sentence construction rules were highly used for indicating purposes of activities and terminologies:

1-1 Raise or lower your knees to adjust for elevation on the target.

1-2 ... supporting fires to allow screening units to break contact and withdraw.
Table 5: Clustering function word list

| Classification | Rank in word list | Frequency | Function words |
|----------------|-------------------|-----------|---------------|
| Group 1        | 3                 | 2667      | to            |
|                | 10                | 889       | for           |
| Group 2        | 4                 | 2348      | of            |
| Group 3        | 7                 | 1298      | in            |
|                | 391               | 46        | under         |
| Group 4        | 9                 | 1007      | figure        |
|                | 49                | 284       | table         |
|                | 24                | 509       | when          |
|                | 25                | 504       | as            |
| Group 5        | 69                | 204       | during        |
|                | 146               | 107       | before        |
|                | 131               | 115       | after         |
|                | 240               | 67        | until         |
| Group 6        | 38                | 377       | if            |
|                | 43                | 328       | should        |
| Group 7        | 45                | 312       | must          |
|                | 64                | 214       | may           |
| Group 8        | 176               | 93        | such          |

1-3 ... with a carry bag, which provides space to carry a CLU, a technical manual ...

1-4 ... improving the kill mechanisms of our missiles to defeat the improved armor.

1-5 ... is used to develop the skills required to engage targets under field conditions.

When “for” is placed in front of nouns or noun phrases, it represents the purposes of objects, actions, and so on. In this case, “for + nouns (NPs)” were used to explain equipment’s functions or purposes of important operating procedures:

1-6 Track gate adjustment for a bunker involves the perceived size of ...

1-7 ... do MGS self-test for battery.

1-8 The gunner should use the NFOV for classification and recognition.

1-9 Inspect the open end of the round for dirt and foreign material.

1-10 The trainers must know the appropriate combat techniques for employing these weapons.

Field manuals can be considered as a type of equipment user guide. They teach readers how to operate systems or component parts and explain the purposes of operating procedures. Thus, “to-infinitive” and “for” are important grammatical rules for developing specific genre.

Group 2. “nouns (noun phrases) + of + nouns (noun phrases)” composed of terminologies or indicates relationships of nouns:

When “of” is placed between noun phrases, the combinations show its relationships of possession, belonging, or connection. It is a kind of a strong supplementary narrative usage to show the relationship between a noun and another. In this case, the researchers noted that many terminologies were connected by “of” and developed to LBs for domain usages. See follows:

2-1 Ensure all the standard principles of camouflage are followed.

2-2 ... the Javelin may not be the weapon of choice in the urban environment ...

2-3 See Table 4-2 for frequency of events as required by DA Pam 350-38 STRAC.
2-4 ... not to fire until given the command of execution.

2-5 ... leader selects a primary position and sector of fire for each weapon.

Group 3. Using “in, under” to describe conditions and situations:

“In, under” are words used to describe some procedures or activities that happen in certain conditions or situations. According to the word list, the researchers found that details in US Army field manuals explain some conditions that may be used to initiate some procedures or some specific functions, as follows:

3-1 ... FTT allows gunner training to be conducted in a field environment ...

3-2 It can be employed in all weather conditions as long as the ...

3-3 ... Javelin’s 2,000-meter range allows flexibility in choosing ambush positions.

3-4 ... be used at any time of day under any weather conditions.

3-5 ... and firing the TOW under NBC and limited visibility conditions ...

Group 4. “Figure and Table” as illustration and data explanation:

The high frequency words such as ‘figure and table’ highlights the importance of “illustration and data explanation” in field manuals. Texts are combined with assistant materials (e.g., photos, graphics, sketches, tables, charts, etc.) in order to show procedures, introduce equipment, and analyse the capabilities of weapon systems in a more precise and detailed manner. Those features hint that, “illustration and data explanation” is critical to making readers understand the abstractive domain knowledge, hence, avoiding misuse of weapon systems.

4-1 Figure A-1 shows the probability of survival for ...

4-2 The eyepiece (Figure 1-14) allows the gunner to see the CLU ...

4-3 Figure 5-2 Javelin command launch unit.

4-4 See Table 3-1, a notional training schedule.

4-5 Table 6-1. Armored vehicle kills.

Group 5. Words for describing operating sequences and timing of uses of weapon systems:

Words such as “when, as, during, before, after, until” are adopted to tell users “when or under what kinds of circumstances, someone will or should do something”. In the typical manual genre, those words are not only used to express tenses but also used to express the important operating sequences of weapons.

5-1 WARNING: When firing the M136 AT4, do not place ...

5-2 As Javelin gunners destroy their targets, leaders should ...

5-3 During combat or field training, TOW crews will ...

5-4 ... activating the seeker before assuming a firing position.

5-5 ... soft targets can normally continue to fight after being attacked by light anti-armor weapons.

Group 6. Conditional clauses:

FM use many conditional clauses to give users scenarios, possible situations, the next steps to take or consequences of actions taken. The common sentences structures identified are: “If something happened, someone should do ...” Examples are as follows:

6-1 If a misfire occurs in combat ...

6-2 If facilities and equipment are not available to ...

6-3 If in a firing position, moves the round ...

6-4 If possible, they should construct reinforced position ...

6-5 If the gunner is not engaging a target ...

Group 7. Giving suggestions, indicating importance and anticipating scenarios:

The researchers found in FMs that “should, must, may” indicated three different levels of authors’ intentions. When describing tactic usages, FMs used “should” to give suggestions to readers. This allows for flexibility and does not put constraint on readers’ tactical approaches. When referring to safety procedures, or safety concerns of weapon systems, FMs used “must” to highlight something that is necessary and nonnegotiable. When
FMs used the word “may”, they gave readers scenarios to foresee situations that may happen. Those messages remind readers of early preparation to avoid occurring surprising and emergency incidents.

7-1 Each position should allow flank fire and have cover and ...

7-2 The Infantry should be able to cover dismounted AAs to ...

7-3 However, trainers and leaders must adopt new safety procedures to ensure ...

7-4 To fire the AT4, the firer must apply firm and steady forward pressure to ...

7-5 The launcher electronics may also be damaged.

Group 8. Using “such as” to give lists of items or examples. See as follows:

e.g. 8-1 Backlighting occurs when an IR source, such as a tank’s exhaust, emits IR ...

e.g. 8-2 ... can also be used against soft targets, such as bunkers, field fortifications, automobiles, and ...

e.g. 8-3 is heat produced by a slow (such as a bonfire) or very quick (such as ...

e.g. 8-4 ... one time on a prearranged signal such as a command, whistle, booby trap, mine, or ...

e.g. 8-5 ... an object in the target scene, such as a far tree line.

(3) Military TISL taxonomy

(i) Checking and refining wordlist and keyword list

Word list and keyword list are an important analysis results from corpus software. However, words such as function words, meaningless words, and some characters existed abundantly on those lists. Thus, the first filtering process is to eliminate those kinds of words for making word list and keyword list more domain oriented.

(ii) Checking LBs of tokens

In this case, the researchers found that some terminologies may exist in the form of phrases. Thus, it is necessary to check LBs of tokens. The researchers based on the setting of cluster size (min.2 and max.5) and term position (both on the left and right) to check each tokens on the keyword list and recheck potential military-oriented words on the wordlist. For example, a the word, “top” is ranked No. 585 in the keyword list may be irrelevant to the military domain if the focus is only on the surface explanation of the word. Nevertheless, when searching for the word in the clusters/n-gram, the results showed “top attack”, “top attack mode”, “top indicator(s)” on the list (Figure 2). This is confirmed by the military experts that the term belongs to one of the most important terms in the Javelin missile-system. To avoid missing critical information, checking LBs of each potential token is crucial.

(iii) Checking concordance lines of abbreviations and acronyms

In the word list and keyword list, the researchers found that FM adopted many acronyms to form terminologies. In addition, the researchers also classified “abbreviations” and “acronyms” as different groups to highlight their importance. Thus, understanding the LBs of acronyms is crucial, otherwise it will be hard to comprehend the terminological meanings. According to US military FM, “abbreviations” and “acronyms” have been explained in detail, but retrieving information directly from those FM seems to be inefficient and lack integration of knowledge. AntConc 3.5.8 (Anthony, 2019) is an appropriate platform for providing concordance evidence to extract LBs of acronyms and abbreviations, and military domain knowledge (see Figure 3).

(iv) Re-categorising military TISL

Keywords and high frequency words represent identity, core knowledge, critical information and specific terminologies of the target corpus. Referred to Munoz’s (2015) research, the experts gathered based on their specialties (1) to eliminate function words, meaningless words and unrelated letters on the word list and keyword list; (2) checking LBs of tokens to avoid missing critical phrase-style terms; (3) checking concordance lines of abbreviations and acronyms to extract definition of those; and (4) classify terminologies into seven groups in order to illustrate the whole frame of military TISL in US Army FM of antitank weapon systems based on terminologies’ functions, meanings, usages, and characteristics. The categorisation (Figure 4) can be defined as: Group 1. Weapon systems; Group 2. Critical component parts and accessories; Group 3. Procedures, actions, and operations; Group 4. People; Group 5. Measurements; Group 6. Abbreviations; Group 7. Acronyms. The groups were created based on aforementioned criteria. The compartmentalisation of military TISL in this case, facilitates the efficiency of understanding the military professional terminologies.
Group 1. The categorisation “Weapon systems” indicated terms which referred to weapons (1-1, 1-2, 1-3) and ammunition (1-4, 1-5):

1-1 The TOW is mainly an antitank weapon used for ...

1-2 The Javelin is a fire-and-forget, shoulder-fired ...

1-3 LAW is a lightweight, self-contained, anti-armor weapon ...

1-4 The encased missile is a priority item and should be ...

1-5 The Javelin round is bulky and restricts movement in heavily ...

Group 2. The categorisation, “Critical component parts and accessories”, represented important parts and accessory equipment, for example, digital controllers (2-1), battery (2-2), critical parts of missiles (2-3), situation indicators (2-4), observation equipment (2-5):
4-1 When the gunner fires the Javelin, it ignites a launch ...

4-2 The TOW gunner or squad leader can use the night sight method or the ... 

4-3 ... identifying targets (to include friend or foe), prioritizing target's ...

4-4... warhead is capable of defeating any known enemy armor. 

4-5 The commander chooses a method of assessing the trainers...

Group 5. The categorisation, “Measurements”, embraced units of measurements (5-1, 5-2, 5-5), ranges of weapons (5-3), and tools of measurements (5-4):

5-1 Russian 40 mm Antitank Grenade Launcher, RPG-7V (Figure H-6).

5-2 Armor penetration: 450 to 2,000 mm.

5-3 The Javelin’s 2,000-meter range allows flexibility in choosing ambush positions.

5-4 Most armies use laser range finders and target designators.

5-5 ... estimate it as a fast-moving vehicle (10 mph or faster).

Group 6. The categorisation, “Abbreviations”, showed the urgency and efficiency of military messages while communicating:

6-1 ... focus adjust (FOC ADJ), sight select (SGT SEL), and filter select (FLTR SEL) switches.

6-2 ... course must be conducted in accordance with (IAW) the Javelin POI established by the US IS.

6-3 ... gate adjust (GATE ADJ), contrast and brightness (CTRS and BRT), and attack select (ATTK SEL) ...

6-4 ... sight select (SGT SEL), and filter select (FLTR SEL) switches.

6-5 The gunner pushes the attack select (ATTK SEL) switch on the right handgrip to ...

Group 7. Finally, the categorisation, “Acronyms”, showed combinations of words developed into terminologies, those may represent weapons (7-1), equipment (7-2, 7-3), and tactical terms (7-4, 7-5):
7-1 **TOW**: tube-launched, optically tracked, wire-guided.

7-2 The day **field-of-view (FOV)** can be used for surveillance and target ...

7-3 **BCU**: battery coolant unit.

7-4 **AAs**: avenues of approaches.

7-5 **OPFOR**: opposing force.

To sum this section, the contributions of the proposed approach can be summarised as follows: (1) results of syntax analysis provide EFL troopers with syntax patterns that high frequently used in the target corpus for facilitating their military information reading and translating efficiency, (2) results of military TISL taxonomy enhance EFL troopers’ TISL acquisition efficiency, and extract technical information in detail. The results presented in this paper indicate insights into the types of syntaxes and TISL used in US Army FMs of anti-tank weapon systems. The analysis made by expert assessment team enabled the results of a corpus-based approach based on linguistic and domain aspects. The findings reveal important pedagogic implications in military training courses at EFL military training facilities where US army anti-tank weapon systems are adopted by them.

### CONCLUSION

Syntax analysis and vocabulary taxonomy are also critical for improving the accuracy and efficiency of corpus analysis and NLP. Military technical information is an uncommon scientific field; if the military simply seek linguists or information engineers’ assistance in processing military information, the analytical results might be distorted especially some information seems insignificant but embedded deeply in domain knowledge.

Language is an important channel to communicate and to acquire information, but it evolves in different domains. Information processing programs based on certain algorithms may not handle complicated NLs’ linguistic rules nor generate high precision resulting data to satisfy each domain. Corpus programs or NLP techniques are ideal toolkits, but machines are not always 100% accurate, thus proper manual annotation is inevitable. The researchers integrated a corpus software, linguists, and domain experts’ specialties to process information, and to make resulting data more meaningful and more applicable to military training purposes.

This paper highlights the value of compiling a narrow-angled specialised corpus to conduct syntax analysis and domain-oriented TISL taxonomy especially customised to address the needs in specific areas with the collaboration of linguists and domain experts, rather than conducting general linguistic analysis. More specifically, this paper suggests that when conducting corpus-based approaches in processing ESP cases, researchers should recruit domain experts to process the linguistic evidence of the specific corpus. In processing the corpus of US Army anti-tank weapon systems FMs, the proposed approach can consolidate and analyse the idiomatic syntaxes from the perspective of a linguist by clustering function words from the wordlist of the target corpus, and categorise military TISL from the perspective of military experts by cross checking wordlist and keyword list, checking LBs to retrieve terminological phrases, and checking concordance lines to retrieve complete terms of acronyms and abbreviations.

The proposed approach highlights the values of combination of a corpus-based approach and related experts’ cooperation in data processing. The significant features can be summarised as follows: (1) the results of syntax analysis and military TISL taxonomy are more in accordance with EFL troopers’ needs in learning military knowledge in English, (2) the proposed approach can integrate large amounts of domain texts and be smoothly utilised by linguist and military experts to conduct in-depth analysis and decipher during information processing, (3) the proposed approach adopts AntConc 3.5.8 (Anthony, 2019), free costs, open access, and with user-friendly operating platforms, to reduce the costs and enhance the efficiency of texts information processing; this approach especially suitable for military that has low defense budget to develop training materials.

In the future, the linguistic analytical results can become valuable reference data and criteria of TISL taxonomy and identification for improving the efficacy of ESP courses developments, and for enhancing the accuracy of corpus analysis to rapidly fetch key information.

### Acknowledgement

The authors would like to thank the Ministry of Science and Technology, Taiwan, for financially supporting this research under Contract No. MOST 109-2410-H-145-001 and MOST 109-2410-H-145-002.

### REFERENCES

Anthony L. (2019). AntConc (Version 3.5.8): Corpus Software. Available at [https://www.laurenceanthony.net/software/antconc/](https://www.laurenceanthony.net/software/antconc/), Accessed 30 April 2020.
Beller S. & Bender A. (2017). Theory, the final frontier? A corpus-based analysis of the role of theory in psychological articles. *Frontiers in Psychology* **8**: 951. DOI: https://doi.org/10.3389/fpsyg.2017.00951

Brindle A. (2016). A corpus analysis of discursive constructions of the Sunflower Student Movement in the English-language Taiwanese press. *Discourse and Society* **27**(1): 3–19. DOI: https://doi.org/10.1177/0957926515605957

Chen L.C. & Chang K.H. (2019). Creating a novel type of information database of military terminology: an example of US Army medical and casualty evacuation. *Basic and Clinical Pharmacology and Toxicology* **125**(S9): 8. DOI: https://doi.org/10.1111/bcpt.13358

Chen L.C., Chang K.H. & Chung H.Y. (2020). A novel statistic-based corpus machine processing approach to refine a big textual data: an ESP case of COVID-19 news reports. *Applied Sciences-Basel* **10**(16): 5505. DOI: https://doi.org/10.3390/app10165505

Cheng K.H. & Tsai C.C. (2019). A case study of immersive virtual field trips in an elementary classroom: Students' learning experience and teacher-student interaction behaviors. *Computers and Education* **140**: 103600. DOI: https://doi.org/10.1016/j.compedu.2019.103600

Cho H. & Yoon H. (2013). A corpus-assisted comparative genre analysis of corporate earnings calls between Korean and native-English speakers. *English for Specific Purposes* **32**(3): 170–185. DOI: https://doi.org/10.1016/j.esp.2013.03.001

Christ A., Penthin M. & Kroner S. (2019). Big data and digital aesthetic, arts, and cultural education: hot spots of current quantitative research. *Social Science Computer Review* **39**(5): 821–843. DOI: https://doi.org/10.1177/0894439319888455

Coats S. (2019). Articulation rate in American English in a corpus of youtube videos. *Language and Speech* **63**(4): 799–831. DOI: https://doi.org/10.1177/00238309198894720

Daskalovska N. (2015). Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning* **28**(2): 130–144. DOI: https://doi.org/10.1080/09588221.2013.803982

Derbentseva N., Safayeni F. & Canas A. J. (2007). Concept maps: Experiments on dynamic thinking. *Journal of Research in Science Teaching* **44**(3): 448–465. DOI: https://doi.org/10.1002/tea.20153

Ferguson G. (2001). If you pop over there: a corpus-based study of conditionals in medical discourse. *English for Specific Purposes* **20**(1): 61–82. DOI: https://doi.org/10.1016/S0889-4

Flowerdew L. (2020). The Academic literacies approach to scholarly writing: a view through the lens of the ESP/Genre approach. *Studies in Higher Education* **45**(3): 579–591. DOI: https://doi.org/10.1080/03075079.2019.1576165

Gonen S.I.K. (2019). A qualitative study on a situated experience of technology integration: reflections from pre-service teachers and students. *Computer Assisted Language Learning* **32**(3): 163–189. DOI: https://doi.org/10.1080/09588221.2018.1552974

Green C. (2019). Enriching the academic wordlist and secondary vocabulary lists with lexicogrammar: toward a pattern grammar of academic vocabulary. *System* **87**: 102158. DOI: https://doi.org/10.1016/j.system.2019.102158

Harvey-Scholes C. (2018). Computer-assisted detection of 90% of EFL student errors. *Computer Assisted Language Learning* **31**(1-2): 144–156. DOI: https://doi.org/10.1080/09588221.2017.1392322

Henry A. & Roseberry R.L. (2001). A narrow-angled corpus analysis of moves and strategies of the genre: 'Letter of application'. *English for Specific Purposes* **20**(2): 153–167. DOI: https://doi.org/10.1016/S0889-4906(99)00037-X

Hsieh J.S.C., Wu W.C.V. & Marek M.W. (2017). Using the flipped classroom to enhance EFL learning. *Computer Assisted Language Learning* **30**(1-2): 1–21. DOI: https://doi.org/10.1080/09588221.2015.1119110

Kim J.E. & Nam H. (2019). How do textual features of L2 argumentative essays differ across proficiency levels? A multidimensional cross-sectional study. *Reading and Writing* **32**(9): 2251–2279. DOI: https://doi.org/10.1007/s11145-019-0947-6

Koops C. & Lohmann A. (2015). A quantitative approach to the grammaticalization of discourse markers evidence from their sequencing behavior. *International Journal of Corpus Linguistics* **20**(2): 232–259. DOI: https://doi.org/10.1075/ijcl20.2.04koo

Lai C., Yeung Y. & Hu J.J. (2016). University student and teacher perceptions of teacher roles in promoting autonomous language learning with technology outside the classroom. *Computer Assisted Language Learning* **29**(4): 703–723. DOI: https://doi.org/10.1080/09588221.2015.1016441

Li L.Q., Zhao J., Hou L., Zhai Y.K., Shi J.M. & Cui F.F. (2019). An attention-based deep learning model for clinical named entity recognition of Chinese electronic medical records. *BMC Medical Informatics and Decision Making* **19**(1): Article Number 235. DOI: https://doi.org/10.1186/s12911-019-0933-6

Li Q., Li S.B., Zhan S., Hu J. & Hu J.J. (2019). A review of text corpus-based tourism big data mining. *Applied Sciences-Basel* **9**(16): Article Number 3300. DOI: https://doi.org/10.3390/app9163300

Li S.L. (2016). A corpus-based study of vague language in legislative texts: strategic use of vague terms. *English for Specific Purposes* **45**: 98–109. DOI: https://doi.org/10.1016/j.esp.2016.10.001

Liu J. & Han L.N. (2015). A corpus-based environmental academic word list building and its validity test. *English for Specific Purposes* **39**: 1–11. DOI: https://doi.org/10.1016/j.esp.2015.03.001

Munoz V.L. (2015). The vocabulary of agriculture semi-popularization articles in English: A corpus-based study. *English for Specific Purposes* **39**: 26–44. DOI: https://doi.org/10.1016/j.esp.2015.04.001

Nekrasova-Beker T.M. (2019). Discipline-specific use of language patterns in engineering: A comparison of published pedagogical materials. *Journal of English for Academic Purposes* **41**: 100774.
DOI: https://doi.org/10.1016/j.jeap.2019.100774
Shanahan T. & Shanahan C. (2017). Disciplinary literacy: Just the FAQs. Educational Leadership 74(5): 18–22.
Sholokhovnn A., Kinnunen T., Vestman V. & Lee K. A. (2020). Voice biometrics security: Extrapolating false alarm rate via hierarchical Bayesian modeling of speaker verification scores. Computer Speech and Language 60: 101024.
DOI: https://doi.org/10.1016/j.csl.2019.101024
Siefritd C., Grosjean J., Lefebvre T., Rollin L., Darmoni S. & Schuers, M. (2020). Evaluation of automatic annotation by a multi-terminological concepts extractor within a corpus of data from family medicine consultations. International Journal of Medical Informatics 133: 104009.
DOI: https://doi.org/10.1016/j.ijmedinf.2019.104009
Smalheiser N.R Luo, M.Q.Addepalli S. & Cui X. K. (2019). A manual corpus of annotated main findings of clinical case reports. Database-The Journal of Biological Databases and Curation 2019: 143.
DOI: https://doi.org/10.1093/database/bay143
Trembach S. (2019). From information to knowledge to wisdom: the cold war battle for information superiority and its implications for thriving in the age of data smog. LIBRI 69(1): 1–12.
DOI: https://doi.org/10.1515/libri-2018-0008
US Department of the Army (2001). Light Antiarmor Weapons: FM 3-23.25. US Department of the Army, Washington DC, USA.
US Department of the Army (2003). TOW weapon system: FM 3-22.34. US Department of the Army, Washington DC, USA.
US Department of the Army (2008). Javelin - close combat missile system, medium: FM 3-22.37. US Department of the Army, Washington DC, USA.
Viswanathan U.M., Sultana A.S. & Shankar S. (2020). Linguistic resources to facilitate science education Current Science 118(2): 271–273.
DOI: https://doi.org/10.18520/cs/v118/i2/271–273
Zygouris-Coe V.I. (2012). Disciplinary literacy and the common core state standards. Topics in Language Disorders 32(1): 35–50.
DOI: https://doi.org/10.1097/TLD.0b013e31824561a