# Designing Computer-Assisted Translation Software for Trainee Translators Based on Classroom Parallel Corpora

**Jaerang Lee**
Ewha Womans University

*The machine translation industry's shift from Statistical Machine Translation (SMT) to Neural Machine Translation (NMT) has recently brought renewed focus on Computer-Assisted Translation (CAT) in translation pedagogy. The purpose of this paper is to propose a CAT tool based on classroom parallel corpora. Trankit, a CAT software prototype the author designed, offers a powerful web-based platform for trainee translators along with a wide range of advanced search options and linguistic information such as concordance lines, syntax trees, and frequency information so that students can make full use of their corpus databases. Section 2 and 3 are literature reviews on parallel corpora and CAT, which are the central concepts forming the theoretical background of the design. Section 4 asserts the necessity of a new CAT tool in relations to these two underlying concepts. Section 5 presents the user interface of the program, exploring applications of CAT in translation training.*

**Keywords: parallel corpora, translation classroom, machine translation, translation pedagogy, Computer-Assisted Translation (CAT)**

## 1. Introduction

As Google Translate's shift from Statistical Machine Translation (SMT) to Neural Machine Translation (NMT) in the late 2016 has led to a significant leap in machine translation quality (Luong, Sutskever, Le, Vinyals and Zaremba 2014; Sennrich, Haddow and Birch 2015; Wu et al. 2016), professional translators are now

faced with greater challenges from their machine competitors. The remarkable improvement of Machine Translation (MT) engines in the recent years prompted many translation schools to incorporate Computer-Assisted Translation (CAT) into their translation curricula. Graduate School of Translation and Interpretation at Ewha Womans University launched the Introduction to CAT course in 2017 in an attempt to train students how to integrate CAT tools in their translation workflow. Graduate School of Interpretation and Translation at Hankuk University of Foreign Studies also provides the CAT course for second-year MA students. School of English Language and Literature at Sookmyung Women's University introduced Localization Program in Fall 2016, offering courses on Localization Tool and Process (Chun 2017).

The increased attention to MT in Translation Studies should be explored in the context of today's translation landscape. Translation practices have grown inseparable from computer technology ranging from programs as simple as word processors to localization tools and Google Translate. In this sense, today's translation can be characterized as a form of human-computer interaction (O'Brien 2012; Olohan 2011).

While much research has been undertaken on the MT era's effects on Translation Studies (Choi 2017; Chun 2017), relatively little attention has been given to the issue of how human translators should use computer technology, particularly CAT, to their own advantage. Previous studies on the design of CAT have concentrated on technical aspects of CAT tools such as sentence alignment and terminology extraction, mostly carried out by computer scientists and machine translation scholars (Barrachina et al. 2009; Dagan and Church 1994; Zajac and Vanni 1997).

The present paper discusses the necessity of a new CAT tool from the perspective of trainee translators. It is based on the idea that today's translation is marked by interactivity between humans and computers, where effective utilization of computer technology largely determines one's translation competence. The key goal of this paper is to put forward Trankit, a CAT software prototype I designed. The computer program offers a web-based platform for trainee translators along with a wide range of advanced search options and linguistic information such as concordance lines, syntax trees, and frequency information.

Trankit lies at the interface of CAT technology and parallel corpora. Section 2 and 3 of this paper are mainly literature reviews on parallel corpora and CAT, respectively. Section 4 asserts the need for a new educational CAT tool in relations to these two underlying concepts, answering the questions: What role parallel corpora play in developing an effective CAT tool and why it is particularly important to have an organized classroom corpus database? The design prototype of the program is presented in Section 5. Ultimately, this study aims to bring into focus the potential of classroom parallel corpora and to explore applications of CAT in translation training.

## 2. Parallel Corpora in Translation

This section primarily consists of literature reviews on parallel corpora in translation, which serve as the basis of the CAT program that will be presented in Section 5. The concept of parallel corpora in translation will be discussed in three different aspects: MT, translators' problem-solving strategies, and translation classrooms.

### *2.1. Parallel Corpora in MT*

Parallel or bilingual corpora are a collection of texts in a source language and their translated counterparts. The rapid development of NMT is largely attributed to the exponential growth of bilingual texts on the web. A plethora of parallel corpora —web-crawled and manually established—have been used as important training materials for improving NMT (Brown, Lai and Mercer 1991; Munteanu and Marcu 2005; Sennrich et al. 2015). A landmark study by Bahdanau, Cho, and Bengio (2014), which contributed to the paradigm shift from SMT to NMT, relies heavily on parallel training corpora. However, when it comes to less widely spoken languages, it is harder to obtain high-quality parallel texts because many of the parallel resources for MT training originate from political documents for international

entities such as European Parliament Proceedings Parallel Corpus (Markantonatou et al. 2006; Munteanu and Marcu 2005). The lack of parallel corpora in low-resource language pairs is pointed to as one of the reasons why Korean-English MT, for instance, generally shows lower performance than in major language pairs like French-English (Chen, Liu, Cheng and Li 2017; He et al. 2016) As a result, there is soaring demand for parallel corpus data in the MT industry. Flitto, a crowdsourced translation platform, earns about 90 percent of its revenue from selling language data to tech giants like Google and Naver (Bischoff 2015).

### *2.2. Parallel Corpora as Problem-solving Strategy*

Parallel corpora are useful resources not only for MT training but also for human translators. If parallel corpora are defined as a pair of source texts (ST) and target texts (TT), they are found everywhere in translation practices. Translators often encounter challenges due to syntactic, lexical and pragmatic differences between two languages. They may spend more than a few minutes on finding a right equivalent for a single word. Obviously, this is a stressful experience for translators. Wilss (1996) argues that obstructions in a decision-making process in translation caused by an excess of alternatives and a delay in information collection may result in "no-choice behavior," a tendency to make hasty and unreasonable decisions without full deliberation. He, therefore, asserts that it is important to simplify this demanding cognitive process.

In this case, parallel corpora can serve as an effective problem-solving strategy in that they can facilitate the search and retrieval of the most appropriate word choice from a translator's cognitive pool. This retrieval process is defined in various ways by researchers who described translation as a problem-solving process (Ordudari 2007). Levý, Althoff, and Vidal (2012) stated that translation is a decision-making process, where one first defines the class of alternatives (i.e., the paradigm) before choosing among them within the context. Tirkkonen-Condit (2000) concluded through her think-aloud experiments that translators come up with a number of tentative solutions to a problem, giving them positive or negative evaluation before

acknowledging TT as a final solution. In short, translation output is the product of decision making among a pool of options. Translators sometimes instantly come up with an answer—which Tirkkonen-Condit referred to as "automatic" solution (2000: 126)—and at other times look up alternatives in digital databases, such as the web, dictionaries, and previous projects before making a decision. The European Master's in Translation (EMT) expert group defines the ability to do the latter as "information mining competence" and views it as a crucial component of translation competence (Gambier 2009). The term includes the capacity for terminological research, mastery of tools and search engines, and archiving. Taking all these into account, it is evident that parallel corpora can play a key role in the retrieval process of translation problem-solving.

## *2.3. Parallel Corpora in Translation Classrooms*

Highlighting the importance of parallel texts, numerous scholars have looked into applications of parallel corpora in translation training. Sharoff (2006) reviewed several case studies on the use of comparable corpora as a problem-solving strategy and established a four-step translation methodology for trainee translators. Zanettin (1998) explored how to train translators with bilingual concordancers such as Wordsmith Tools and Paraconc. Some scholars focused on the effect of translation curricula using parallel corpora. Gallego-Hernández (2015) conducted a survey-based study on English-Spanish translators to see their use of corpora from various aspects: Which types of corpora they use, when they build DIY corpora and how useful they are, etc. Frankenberg-Garcia (2015) conducted an experiment where she asked her M.A. students in Translation at the University of Surrey to build DIY corpora with or without corpus-building tools including the WebBootCaT tool and assessed the trainees' reaction about the new teaching method at the end of the semester. The study concluded that corpora were not only used to find more appropriate translation options but also to analyze collocations and check frequencies, which conventional search methods could not have so easily provided.

Despite the attention given to the application of parallel texts to translation

curricula, Translation Studies has rarely focused on the characteristics of corpora produced by trainee translators. The classroom parallel corpora raise pertinent questions as to how students store and manage them, in which file format they save them, and how distinctive the classroom corpora are in comparison to parallel corpora for MT training.

Parallel corpora in translation classrooms have some distinctive features. First, each ST corresponds to multiple translated counterparts produced by each class participant; in other words, there is more than one TT corresponding to each ST. This is usually not the case with many parallel corpus data for MT and language learner corpora. Second, they are mostly produced as part of assignments subject to grading, which is likely to enhance their overall quality in comparison with web-crawled corpora and random search results. Third, classroom corpora vary in topics as translation schools have extensive curricula that consist of literary and technical translation spanning various genres and themes, unlike many free-use corpora whose topics are often limited to political speeches (Munteanu and Marcu, 2005) or open-source parallel corpora such as OPUS mainly used for machine training (Tiedemann 2012). Given the three distinguishing features of parallel corpora in translation classrooms—one-to-many correspondence, high quality, and variety of topics—it is necessary to devise an electronic system to properly manage and sustain them as academic and pedagogical resources.

## 3. Computer-Assisted Translation Tools for Trainee Translators

### *3.1. What is Computer-Assisted Translation?*

Before exploring the interface between CAT and classroom parallel corpora, the origin of CAT should be first discussed. It is difficult to touch on the history of CAT without mentioning MT since both of them have developed in line with Natural Language Processing (NLP) technology. NLP is a field of computer

engineering, which examines how to process human languages into computer-interpretable versions. NLP technologies such as segmentation, text alignment, and automated terminology extraction have played a pivotal role in the development of both CAT and MT (Barrachina et al. 2009; Bowker and Fisher 2010; Isabelle and Church 1997).

The recent focus on CAT in the translation training scene is not an entirely new phenomenon. The early 1990s saw the advent of the four pioneering CAT programs including Translator's Workbench from Trados, TranslationManager/2 from the IBM Corporation, the Transit system from STAR AG, and Eurolang Optimizer (Hutchins 1998). Since then, CAT has constantly been noted by computer scientists and translation scholars like Martin Kay and Alan Melby.

The discussion in the 1990s on Machine-Aided Human Translation (MAHT), as opposed to Human-Assisted Machine Translation (HAMT), has implications for how we understand the relationship between humans and computers (Zajac and Vanni 1997). Martin Kay was one of the innovators who emphasized the importance of CAT tools after MT's forbidding flaws had been recognized in the 1960s (Bowker and Fisher 2010). His article, "The Proper Place of Men and Machines in Language Translation" originally published in 1980, proposed the translator's workbench approach (Kay 1997), representing the MAHT view. Alan Melby (as cited in Hutchins, 1998) proposed a multi-level translation assistance tool where translators can access various functions, including concordancing and terminology databank, on a single platform. Both scholars emphasized that translators should be in control, defying the traditional notion of HAMT, which reduces human translators to MT's post-editors (Isabelle and Church 1997).

What distinguishes CAT or MAHT from HAMT is that the former regards machines as assistants for human translators, which Kay (1980) described as "the translator's amanuensis." It is true that the recent quality leap in NMT is invoking fear that the machine might take over the translation industry shortly (Bennett and Gerber, 2003; Bundgaard 2017). However, this does not change the fact that at least now computers are playing an auxiliary role in most translation scenes. Moreover, human translators' ability to use computer technology to their own advantage is

becoming more important than ever; CAT tools are at the core of such technologies.

Today's translation is characterized as a form of human-computer interaction where translators widely use word processors, the Google search engine, CAT programs, and even Google Translate. Therefore, the notion of translation competence should be expanded to translation tool skills (Shin 2007). CAT can include, in a broader sense, all kinds of computer tools that assist translators, but it mostly refers to translator's workstations such as SDL Trados Studio designed to facilitate translation tasks (Bowker and Fisher 2010). Translator's workstations have two major components: Translation Memory (TM) and terminology tools. TM is a digital warehouse where users' previous translations are saved for later retrieval. Terminology tools allow users to store terminological information including keywords, equivalents, context, and sources (Goldsmith 2017; Terminology Management 2014).

### *3.2. Computer-Assisted Translation in Translation Classrooms*

This subsection explores how CAT is incorporated into translation classrooms. The theme is currently highlighted in Translation Studies. Many papers emphasize translators' ability to adapt to the fast-changing digital environment. Shin (2017) redefined translation competence in a technological paradigm while O'Brien (2012) viewed translation as a form of human-computer interaction. According to the EMT expert group, translation competence includes "information mining competence" and "technological competence" (Gambier 2009); the ability to use CAT programs or concordance software is at the intersection of both qualities.

As to methods, scholars have taken varied approaches which lie somewhere between HAMT and MAHT (i.e., between adapting to the new MT landscape and taking advantage of advanced computer technology). Depending on which side they belong to, pedagogical methodologies differ. The HAMT view emphasizes post-editing and MT evaluation capabilities while the MAHT view underlines the use of CAT tools.

Regarding the HAMT approach, several researchers experimented on the

pedagogical effect of post-editing courses for trainee translators. Şahin (2014) analyzed how trainee translators reacted to an experimental course where one group was asked to post-edit MT output while the other translated from scratch. Sycz-Opoń and Gałuskina (2017) concluded from their experiment that the post-editing of the MT raw output requires trainee translators of special skill sets distinguished from traditional ones. Another investigation was conducted by Koponen (2015). She shared the result of a newly-introduced post-editing course at Helsinki University in Fall 2014, whose topics stretched from theory and history of MT and post-editing; post-editing without source text; post-editing quality levels and guidelines; MT quality evaluation.

Besides post-editing, some scholars identified the evaluation of MT systems as another central part of translator training. Somers (2001) asserted the need to teach trainee translators about MT and to provide hands-on experiences of related software. He encouraged students to do exercises in post-editing and commenting and to formulate post-editing guidelines so that the trainee translators can familiarize themselves with MT. Kim (2017) asked 30 undergraduate students taking Korean-English translation class to do their assignments with the help of MT and observed their pre- and post-editing practices. Hartley and Schubert (1998) adopted new evaluation criteria in their translation curricula to show how MT can be integrated into translation workflow. They devised several workflow scenarios that involve the evaluation of a variety of MT systems from different perspectives including feasibility, usability and its cost-effectiveness.

On the other hand, there has been relatively little research on the MAHT approach, which views CAT as a pedagogical tool. Kim (2016) conducted case studies on the syllabi of 10 international translation schools to outline three ideal curricula for CAT-translation: CAT Theory, which aims to enhance students' basic understanding of MT and CAT; Introduction to CAT; Applied CAT on how to utilize CAT tools. Yang, Ciobanu, Reiss, and Secară (2017) proposed to use CAT tools for translation quality assessment in translation classrooms in order to encourage students to experience how the translation industry works and establish reliable guidelines on quality evaluation.

## 4. The Necessity of Special CAT Tool for Trainee Translators

Until this point, the concepts of parallel corpora and CAT were discussed individually in Section 2 and 3. This section points out why there should be a special CAT tool based on parallel corpora for trainee translators. One may wonder why we should not just use SDL Trados Studio in CAT curricula since most CAT courses in translation schools in Korea focus on how to incorporate the translation environment program in students' workflow (Chun 2017; Kim 2017). Therefore, a thorough examination should be made on the problem of using the translator's workstation as an educational tool.

SDL Trados Studio, one of the most representative CAT tools utilized by more than 225,000 international users, offers a workbench with a variety of functions including TM, Autosuggest, Autocompletion, and MultiTerm (SDL Trados Studio 2017). However, there are some obstacles that hinder trainee translators from using it as an effective work assistant. The problems are as follows: (1) TM with term-to-term and sentence-to-sentence alignment, (2) system overloads, and (3) difficulty in locating data.

Concerning TM, the limitations of CAT tools have been reported by many professional translators. Bundgaard's (2017) workplace research collected translators' opinions of favor and disfavor towards MT-assisted TM. SDL Trados Studio 2011 was used in the experiment. The study showed the cons were more prominent than the pros especially due to the uselessness of TM in finding matches. The pros, on the other hand, answered that the concordance function was useful in locating specific phrases or words. The result demonstrates that the usefulness of TM hinges upon how accurate the matches in phrases or words are. TM retrieves information after checking whether there is a match between a sentence in question and other previous translations. The decision is made based on the similarity of segments consisting of terms or phrases. Depending on how close they are, the matches of segments are presented in SDL Trados Studio either as "100% Match" or "Fuzzy Match," a match that is less than 100 percent (Goldsmith 2017).

However, this rule of one-to-one similarity may not perfectly work in

Korean-English translation, which is likely to consist of plenty of noisy parallel corpora, as opposed to sentence-aligned parallel corpora where both ST and TT are often typologically close languages (Fung and McKeown 1997). When it comes to Korean-English translation, most bilingual corpora are not composed of texts translated term-to-term, which mean they may not correspond well when segmented by sentence. Many words and phrases are sometimes converted into entirely different translated counterparts, so TM based on matching has a limited effect on solving challenging translation questions. The two source texts below are the examples that show such difficulties. They are a part of a real classroom assignment given to 27 MA students taking the Introduction to Translation and Interpretation course in Spring 2017 at Graduate School of Translation and Interpretation of Ewha Womans University.

ST 1) Yet it's the only way to explain what we've done to the night: We've engineered it to receive us by filling it with light. [1)]

ST 2) Its benefits come with consequences—called light pollution—whose effects scientists are only now beginning to study.

ST 1 shows the difficulty of one-to-one equivalence between bilingual segments. It is particularly tricky to translate the verb "engineer" because it will be an awkward word if rendered in Korean according to the denotative definition: "to manufacture" or "to design" which corresponds to "제작하다" or "설계하다" in Korean. Additionally, the segmentation rule often causes syntactic problems in Korean-English translation where a source sentence is often split into more than one target sentence, or several sentences merged into one. According to the default rule of segmentation in SDL Trados Studio, a segment ends with a full stop (Goldsmith 2017), but it does not apply always. Consider ST 2 for example. Because of the phrase "called light pollution" set apart by em dashes, some translators may choose to break the sentence into two while others may create a complex sentence with a

1) The two ST examples are excerpted from the National Geographic article, "Our Vanishing Nights" (Klinkenborg 2008).

subordinate clause.

In addition to the limitations of parallel alignment, another problem with the use of TM is that students may have to load unnecessarily large data sets to use them as references, which in turn leads to an excessively long response time and a system overload. These frequently occur in many TM systems including SDL Trados Studio, causing dissatisfactions among translators (O'Brien and Moorkens 2014). Due to a large quantity of RAM assigned to TM and its adjunct functions, some experts recommend aligning past data only if they are relevant to the current project that one is working on (Goldsmith 2017). Therefore, an ideal CAT tool for trainee translators should streamline optional functions so that students can locate as many previous data as possible.

In summary, an effective CAT tool for trainee translators should cover a wide range of issues from term-to-term translation to syntactic and organizational problems, be lightweight, and enable quick and thorough search.

## 5. Basic Features of Trankit

This part of the paper puts forward Trankit, a CAT software prototype that I designed, which may serve as an alternative training tool to current CAT programs. Trankit provides a web-based platform for classroom parallel corpora where translation students can facilitate the problem-solving process and gain access to more parallel language data produced by themselves and peer students. Each following subsection will present the five basic features of Trankit.

### *5.1. Cloud-Based File Management*

First, Trankit offers a cloud-based file storage and sharing system. It allows students to have exclusive access to their classroom corpora by means of user registration. To join the system, users should first register and verify their affiliation using the email verification code sent to their institutional email address. Figure 1 illustrates the welcoming page of Trankit. After users sign in, Trankit walks them through the page where multiple course participants share their assignments. The list of classes they take is displayed in the left column under the user name. If they want to access their own corpora only, they can view their database history by entering the My Corpora menu. After clicking Start Browsing, they can search and analyze parallel language data and compare theirs with others.

One of the great advantages of web-based corpora management tool is file-sharing among peers. When students face challenges while translating, the platform allows them to refer to their past problem-solving experiences or exemplary translations of their peers. Without the web-based platform, it would be difficult for students to pinpoint which digital file, among a stack of assignment files on their local computers, contains the word, sentence, or grammatical rule in question.

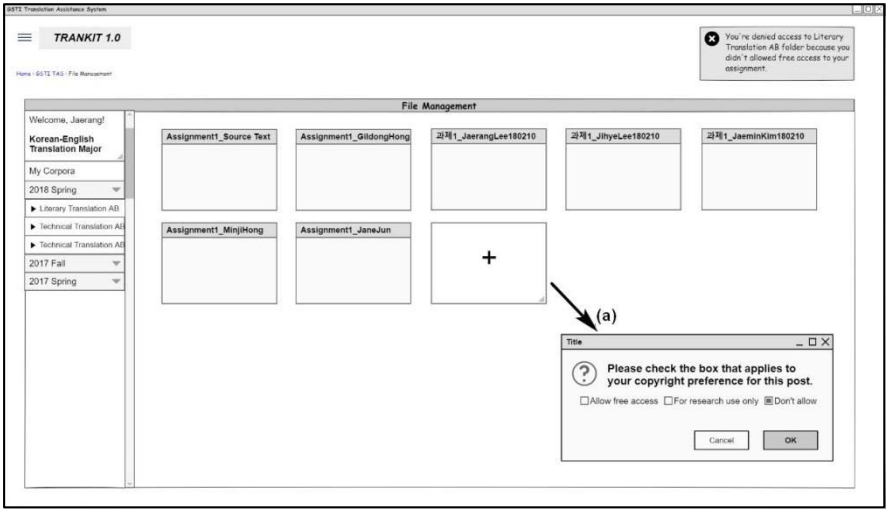


Figure 1. Trankit as a cloud-based file-sharing platform

Copyright in translation is a thorny issue. The notice box to which the arrow (a) in Figure 1 points allows users to choose among copyright options when sharing their post. It is an essential function for individuals and translation institutions that attempt to manage classroom corpora as their permanent educational or research resources. In Trankit, users can choose whether they will allow free access, permit access only for research use, or deny all access. The default option is to allow free use, but they can restrict other students' access by selecting "don't allow." However, if this one is selected, they will also be prohibited from accessing others' translations.

### *5.2. All-in-one interface*

Figure 2 shows the central interface of Trankit, which consists of the File Management Tool in the left column, the Concordancer in the center, the Directory, and the Browsing History Tool in the right. Students can access this interface either by selecting files in the welcoming page illustrated in Figure 1 or by loading them manually. If users want to load all their previous projects, manual loading would be preferred. After .txt files or folders are loaded, the number of the total words contained in the archives will appear at the bottom of the File Management menu. The Convert to .txt Button in the left column provides a link to Multi-Doc File Converter, which converts .docx files in bulk to .txt files, the only readable file extension for the concordance software. Using Multi-doc File Converter, students can upload the entire folder in which their multiple assignments are saved and convert them to .txt files only within a few seconds. The all-in-one process helps students reduce the time to search for the information they need in their messy directories.

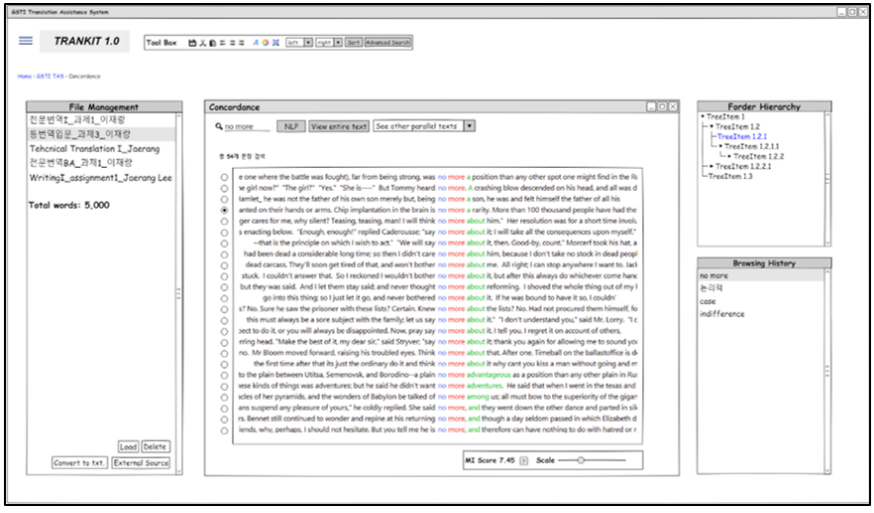


Figure 2. The main page of Trankit

Uploaded files are not automatically sentence-aligned, as are in other CAT programs like SDL Trados Studio. Instead, users can access the corpus counterpart by simply clicking a sentence, and then the screen will display the entire ST alongside the TT. It is due to the frequent errors in segmentation and sentence alignment in noisy parallel corpora as discussed in Section 4. This file alignment function, as opposed to sentence alignment, can be useful for students who store ST and TT in separate files. In many translation classes, if an instructor uploads an assignment file with ST, students submit the TT saved in a separate file. Thus, some students will need to find the counterpart using the file-alignment system.

### *5.3. Concordancer*

A concordancer is a tool which creates concordance lines which include a keyword and other sentence components embracing it left and right. As shown in Figure 2, a user first inputs a keyword in the Search Bar on the upper left side of the central column. Then, the search result will show the sentences that include the keyword and will activate the selection boxes on the left for further linguistic and

statistical analysis.

With the help of the Concordancer, users can find out in what context the search word was used, how frequently it appears, and what other words it most often collocates with. On concordance lines, the three words to the left or right of the key term can be highlighted in different colors for easy recognition. Users can adjust the drop-down menu in the Tool Bar at the top of the screen to decide how many words to be highlighted to the left and the right. They also can use the Sort button to rearrange it in alphabetical or reverse alphabetical order. When they click the keyword, the screen will show the entire text alongside its translated counterpart. By using the drop-down menu that reads "See other parallel texts" at the top of the Concordancer, students can access their peers' files so that they can immediately compare among the expressions used in other translations.

The Concordancer is especially useful when students struggle with translating expressions unique to English. Trankit allows users to refer to their past problem-solving experiences by creating concordance lines. For instance, if they input em dash in the Search Bar, Trankit combs through their individual and peer corpora and shows every sentence that includes the punctuation mark. Here we revisit ST 1 mentioned in Section 4 and take a look at how the class attendants took different translation strategies in their target texts.

| | |
|---|---|
| ST 1 | Yet it's the only way to explain what we've done to the night: We've engineered it to receive us by filling it with light. |
| TT1(a) | 우리는 공학의 힘을 동원해 밤을 빛으로 채움으로써 어둠을 정복해 왔다. |
| TT1(b) | 우리는 밤이 빛으로 가득 차도록 조명을 설계해 밤이 우리를 수용하도록 했다. |
| TT1(c) | 우리는 어둠 속에 빛을 채워 밤이 우리 인간을 받아들이도록 했다. (∅) |

The above examples show how the concordance function can help students retrieve many different options for the translation of the verb "engineer." Both TT1(a) and TT1(b) reveal the word's denotative meaning, but with different expressions. On the other hand, the translator of TT1(c) omitted it to make the

translation simpler. By sharing their corpora on the platform, students can later refer to other possible translation options suggested by peer students.

### *5.4. Statistical Analysis*

Trankit displays the frequency data of how many times a key term appears in a set of uploaded files. This function is useful for translation students who need to check, for example, whether the noun "proof" collocates more often with the adjective "solid" or with "concrete." The frequency information is provided in terms of Mutual Information Score (MI Score) after a simple calculation and is displayed at the bottom of the Concordancer. It demonstrates the possibility of two words' collocating with each other within their overall frequency. The MI level is calculated in the same way as in BYU corpora, which is originally suggested by Church and Hanks (1990):

> MI Score = log2{(AB × Corpus size) / (A × B)}
> (A= frequency of key term, B= frequency of collocate, AB = the co-occurrence of A and B)

This statistical information is especially effective when a user loads a large-scale corpus database because the MI Score, according to the formula, increases in line with the corpus size. The Scale Widget indicates the MI level visually on the scale, which allows users to choose the best option among many collocation possibilities.

### *5.5. Linguistic Analysis Using Natural Language Toolkit for Python*

Trankit also offers a wide range of advanced search options and linguistic information such as concordance lines, syntax trees, and frequency information. Natural Language Toolkit for Python (NLTK) provides open-source text processing libraries for NLP programming in Python language. NLTK libraries offer modules for classification, tokenization, stemming, tagging, parsing, and semantic reasoning (Bird, Klein and Loper 2009).

Table 1. The key features of NLTK

| Modules | Functions |
|---|---|
| Collocations | Collocation finder |
| Grammar | Drawing parse tree |
| Probability | Frequency distribution, conditional probability |
| Text | Concordancing, count, dispersion plot |
| Tree | Syntax tree and morphological tree |

If users need further linguistic analysis of a sentence, they can access NLTK by selecting the sentence before clicking the NLP Button next to the Search Bar. The key modules of NLTK include collocations, grammar, probability, text, and tree, each of which will help students dissect a sentence to get the full understanding of it. The toolkits can also tokenize, parse, and tag a sentence, allowing students to analyze a complex sentence structure as illustrated in Figure 3.

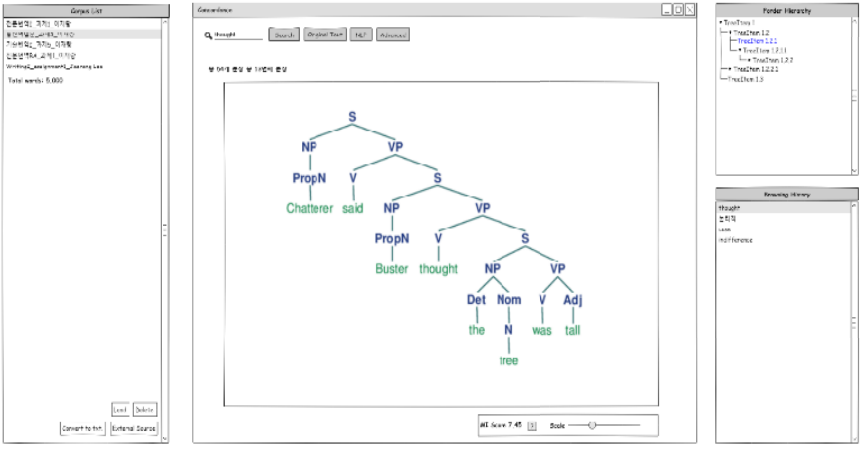


Figure 3. Screenshot of Parse Tree Drawn Using NLTK

## 6. Conclusions

In this paper, the importance of parallel corpora and CAT was discussed through the literature reviews in Section 2 and 3. Section 4 asserted the need for a new CAT tool for trainee translators by stating several limitations of TM-based translation workstations, mainly of SDL Trados Studio. Finally, the prototypical design of Trankit was presented in Section 5.

Trankit sets itself apart from other CAT tools in that it is a lightweight, web-based file-sharing platform. First, it allows for organized management and sharing of classroom parallel corpora. The shared data can also be used by translation schools for research and pedagogical purposes. Second, Trankit encourages students to archive language data for later retrieval. It helps trainee translators quickly refer to previous works of themselves and peers. Third, it goes beyond terminology-limited search to overall data reference including collocations, syntax and paragraph organization.

This paper proposes a CAT tool not from the perspective of computer engineers but from trainee translators, distinguished from other studies focusing on improving the accuracy of NLP mechanisms in CAT tools. The prototypical design of Trankit displayed with some in-class examples visualizes how the tool can be actually used in translation classrooms. The present study also echoes the need for a change in translation curricula that many scholars have pointed out. Kiraly (2003) emphasized the collaborative, process-based learning as the translation landscape is witnessing a rapid change due to globalization and the 4th industrialization. Other scholars argued the necessity of MT curricula which aim to enhance students' basic understanding of MT, CAT, and Natural Language Processing, as noted in Section 3.2.

If applied to translation classrooms, Trankit can be used in class activities where multiple students share their corpora to form a massive high-quality database. The software will become more powerful if students in the same class share their versions of translations, which can vary significantly in terms of word choice, sentence structure, and tone. Students can examine the differences between their translations by using many advanced functions in Trankit and further apply the

information to their own writings. In addition to the educational benefits, Trankit can contribute to the practical use of parallel corpora that belong to translation schools. Using Trankit, the institutions may use classroom corpora for research and pedagogical purposes after obtaining copyright permission on the platform.

The definition of translation competence is rapidly changing as digital technology has grown inseparable from translation practices. The qualities include not only traditional ones such as language skills but also the ability to manage digital tools and extract information on the web. In this sense, the systematic management of parallel corpora on Trankit will enable trainee translators and translation schools to be better prepared for the change in translators' virtues.

## References


Bahdanau, D., Cho, K. and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. Paper presented at *ICLR 2015*, San Diego.

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vilar, J. and Vidal, E. (2009). Statistical approaches to computer-assisted translation. *Computational Linguistics* 35(1): 3-28.

Bennett, S. and Gerber, L. (2003). Inside commercial machine translation. In H. Somers (Ed.), *Computers and Translation: A translator's guide*. Amsterdam/Philadelphia: John Benjamins, 175-190.

Bird, S., Klein, E. and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. California: O'Reilly Media, Inc.

Bischoff, P. (2015, January 31). Korean entrepreneur went from translating K-Pop tweets to selling language data to web giants. Retrieved from https://www.techinasia.com/korean-entrepreneur-translating-kpop-tweets-selling-language-data-google/ on 26 February 2018.

Bowker, L., Fisher, D. and Van Doorslaer, L. (2010). Computer-aided translation. In Y. Gambier (Ed.), *Handbook of Translation Studies*. Amsterdam/Philadelphia: John Benjamin, 60-65.

Brown, P. F., Lai, J. C. and Mercer, R. L. (1991). Aligning sentences in parallel corpora. The 29th Annual Meeting on Association for Computational Linguistics Proceedings.

Bundgaard, K. (2017). Translator attitudes towards translator-computer interaction-findings from a workplace study. *HERMES-Journal of Language and Communication in Business* 56: 125-144.

Chen, Y., Liu, Y., Cheng, Y. and Li, V. O. (2017, July). A teacher-student framework for zero-resource neural machine translation. Paper presented at *the 55th Annual Meeting of the Association for Computational Linguistics,* Vancouver.

Choi, S. (2017). The technological turn in translation studies: The impact of AI on audiovisual translation. *The Journal of Translation Studies* 18(2): 207-228.

Chun, H. (2017). The 4th Industrial Revolution and the Status of Korean Translation Industry, and the Future of Interpretation and Translation Education. *The Journal of Interpretation and Translation Education* 15(3): 235-261.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1): 22-29.

Dagan, I. and Church, K. (1994). Termight: Identifying and translating technical terminology, The Fourth Conference on Applied Natural Language Processing Proceedings.

Frankenberg-Garcia, A. (2015). Training translators to use corpora hands-on: Challenges and reactions by a group of thirteen students at a UK university. *Corpora* 10(3): 351-380.

Fung, P. and McKeown, K. (1997). A technical word-and term-translation aid using noisy parallel corpora across language groups. *Machine Translation* 12(1): 53-87.

Gallego-Hernández, D. (2015). The use of corpora as translation resources: A study based on a survey of Spanish professional translators. *Perspectives* 23(3): 375-391.

Gambier, Y. (2009). Competences for professional translators, experts in multilingual and multimedi a communication. Retrieved from https://ec.europa.eu/info/sites/info/files/emt_competences_ translators_en.pdf/ on 12 February 2018.

Goldsmith, E. (2017). An introduction to translation memory. Retrieved from https://www.sdltrados.co m/download/an-introduction-to-translation-memory/101902/ on 1 February 2018.

Hartley, T. and Schubert, K. (1998). From testbench to workflow: Relocating MT in education and training. Translating and the Computer 20: Proceedings of Aslib Conference.

He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T. and Ma, W. (2016, December). Dual learning for machine translation. Paper presented at the Advances in Neural Information Processing Systems, Barcelona.

Hutchins, J. (1998). The origins of the translator's workstation. *Machine Translation* 13(4): 287-307.

Isabelle, P. and Church, K. (1997). Preface. *Machine Translation* 12(1): 1-2.

Kay, M. (1997). The proper place of men and machines in language translation. *Machine Translation* 12(1-2): 3-23.

Kim, J. (2016). A Study for the Language-based Convergence Class through the Case Analysis of Foreign Universities. *The Journal of Interpretation and Translation Education* 14(3): 25-43.

Kim, S. (2017). Utilization of MT in Translation Classroom. *The Journal of Interpretation and Translation Education* 15(3): 5-37.

Kiraly, D. C. (2003). A passing fad or the promise of a paradigm shift in translator education? In G. S. Koby and B. J. Baer (Ed.), *Beyond the Ivory Tower: Rethinking Translation Pedagogy*. Amsterdam/Philadelphia: John Benjamins, 3-32.

Klinkenborg, V. (2008, November). Our Vanishing Nights. Retrieved from http://ngm.nationalgeogra phic.com/2008/11/light-pollution/klinkenborg-text on 25 February 2018.

Koponen, M. (2015). How to teach machine translation post-editing? experiences from a post-editing course. Proceedings of the 4th Workshop on Post-Editing Technology and Practice.

Levý, J., Althoff, G. and Vidal, C. (2012). Translation as a decision process. *Scientia Traductionis* 11(1): 72-96.

Luong, M., Sutskever, I., Le, Q. V., Vinyals, O. and Zaremba, W. (2014). Addressing the rare word problem in neural machine translation. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics.

Markantonatou, S., Sofianopoulos, S., Spilioti, V., Tambouratzis, G., Vassiliou, M. and Yannoutsou, O. (2006). Using patterns for machine translation (MT). Proceedings of the 11th Annual Conference of the European Association for Machine Translation.

Munteanu, D. S. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics* 31(4): 477-504.

O'Brien, S. (2012). Translation as human–computer interaction. *Translation Spaces* 1(1): 101-122.

O'Brien, S. and Moorkens, J. (2014, August). Towards intelligent post-editing interfaces. Paper presented at XXth FIT World Congress 2014, Berlin.

Olohan, M. (2011). Translators and translation technology: The dance of agency. *Translation Studies* 4(3): 342-357.

Ordudari, M. (2007). Translation procedures, strategies and methods. *Translation Journal* 11(3): 8.

Şahin, M. (2014). Using MT post-editing for translator training. Retrieved from http://lodel.irevues.inist.fr/tralogy/index.php?id=255&format=print on 27 January 2018.

SDL (2014). Terminology Management (online) Retrieved from https://www.sdltrados.com/download/managing-terminology-with-sdl-multiterm/71998 on 15 February 2018.

SDL (2017). SDL Trados Studio 2017 Professional product brief (online) Retrieved from https://www.sdl.com/download/sdl-trados-studio-2017-professional/110101/ on 22 January 2018.

Sennrich, R., Haddow, B. and Birch, A. (2015). Improving neural machine translation models with monolingual data, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.

Sharoff, S. (2006, May). Translation as problem solving: Uses of comparable corpora. Paper presented at the Third International Workshop on Language Resources for Translation Work, Research & Training, Genoa.

Shin, J. (2007). Developing translator competence as translation tool users. *Conference Interpretation and Translation* 9(1): 111-127.

Shin, J. (2017). Revisiting Translation Competence in a Technological Paradigm. *Interpreting and Translation Studies* 21(4): 51-71.

Somers, H. (2001, September). Three perspectives on MT in the classroom. Paper presented at the MT Summit VIII Workshop on Teaching Machine Translation, Santiago De Compostela.

Sycz-Opoń, J. and Gałuskina, K. (2017). Machine translation in the hands of trainee translators–an empirical study. *Studies in Logic, Grammar and Rhetoric* 49(1): 195-212.

Tiedemann, J. (2012, May). Parallel data, tools and interfaces in OPUS. Paper presented at Lrec 2012, Istanbul.

Tirkkonen-Condit, S. (2000). Uncertainty in Translation Processes. In S. Tirkkonen-Condit and R. Jääskeläinen (Eds.), *Tapping and Mapping the Processes of Translation and Interpreting*. Amsterdam/Philadelphia: John Benjamins, 123-142.

Wilss, W. (1996). Translation as Decision-Making and Choice. In *Knowledge and Skills in Translator Behavior*. Amsterdam/Philadelphia: John Benjamins, 174-191.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M. and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. Retrieved from https://arxiv.org/pdf/1609.08144.pdf/ on 15 January 2018.

Yang, J., Ciobanu, D., Reiss, C. and Secară, A. (2017). Using computer assisted translation tools' translation quality assessment functionalities to assess students' translations. *The Language Scholar* 1(1): 90-105.

Zajac, R. and Vanni, M. (1997). Glossary-based MT engines in a multilingual analyst's workstation architecture. *Machine Translation* 12(1-2): 131-151.

Zanettin, F. (1998). Bilingual comparable corpora and the training of translators. *Meta: Journal Des*

This paper was received on 28 February 2018; revised on 11 May 2018; and accepted on 30 May 2018.

***Author's email address***
lee.jrang1@gmail.com

***About the author***
Jaerang Lee is an MA student in Korean-English Translation at the Graduate School of Translation and Interpretation of Ewha Womans University. She presented at the pre-conference session of the 2017 GSTI academic conference, "Translation & Interpreting: Perspectives on Agency and Role," regarding translator-friendly assistance software based on student corpora.