Tagging a Japanese Learner Corpus of English and Comparing Trigrams with Those in a Corpus of British Students’ Essays

Yoshihito Kamakura

要 旨

本研究は、日本人英語学習者とイギリス人学生の作文から構成されたコーパスを数量的と質的に分析している。その分析において、３語の連なりから成るtrigramの頻度とその構成をイギリス人ネイティブ学生のもと比較し、日本人英語学習者特有の句構成（phraseology）を見出すことを目的としている。さらに、学習者コーパスへのタグ付けのため、二つの異なるタグ付記プログラムを用い、その特性と限界を調べている。

本研究の目的は、日本人英語学習者が多用する句構成を記述することである。英語母語話者とは異なる言い回しが中間言語におけるforeign-likenessと捉えられる。ネイティブの規範と異なる言い回しは、意味の伝達において誤解を呼ぶこともありうる。学習者コーパスに見られる過剰使用（overused）の句構成と通常より頻度が低い（underused）句構成を調べることで、学習者言語では句構成が固定されていることが明らかになった。学習者に共通する句構成を明らかにすることで、柔軟な句構成を用い、より豊かな表現をするよう、教員が指導できる。

本研究の結果として、綴りの間違いがあるものの、上記二つのタグ付記プログラムは学習者の句構成を分析することが十分可能であった。その結果、日本人英語学習者はinを含む前置詞句を文末に使う傾向があり、さらにto不定詞+English+inという句の過剰使用が見られた。一方、受動態の文でinが使用されることが少なく、学習者の過去分詞を用いた句の使用が限られていることを示している。
Keywords: learner corpus (学習者コーパス), quantitative and qualitative features (量的と質的な特徴), trigram (3連語句), phraseology (フレイジオロジー, 句構成), annotation (タグ付け)

Introduction

Learners of a second language may retain unique features in their production, unlike those of the vernacular users of the target language. Learner corpora have been established with the aim of discovering the key attributes of learner languages by contrasting them with those in a corpus of native production. However, the comparison of concordance lines as raw data may be arduous, in particular when analysing the environment of a given word—in this case, the words preceding and following a preposition. A first attempt was made by means of trigrams to consider the adjacent words to the preposition in the corpora of Japanese learners of English and British students as follows:

Table 1  Top ten clusters of multi-word unit with *in*

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Freq.</th>
<th>Cluster</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN THE WORLD</td>
<td>68</td>
<td>IN ORDER TO</td>
<td>19</td>
</tr>
<tr>
<td>IN THE FUTURE</td>
<td>52</td>
<td>IN THE UK</td>
<td>18</td>
</tr>
<tr>
<td>LANGUAGE IN THE</td>
<td>27</td>
<td>IN THE WORLD</td>
<td>17</td>
</tr>
<tr>
<td>IN ORDER TO</td>
<td>22</td>
<td>IN MY OPINION</td>
<td>16</td>
</tr>
<tr>
<td>ENGLISH IN THE</td>
<td>18</td>
<td>AN INCREASE IN</td>
<td>16</td>
</tr>
<tr>
<td>IN ENGLISH AND</td>
<td>16</td>
<td>THE UK</td>
<td>14</td>
</tr>
<tr>
<td>STUDY ENGLISH IN</td>
<td>16</td>
<td>IN THIS COUNTRY</td>
<td>13</td>
</tr>
<tr>
<td>PEOPLE IN THE</td>
<td>16</td>
<td>IN THE U</td>
<td>13</td>
</tr>
<tr>
<td>ENGLISH IN JAPAN</td>
<td>14</td>
<td>IN THE CASE</td>
<td>13</td>
</tr>
<tr>
<td>TO SPEAK ENGLISH</td>
<td>13</td>
<td>THE FIELD OF</td>
<td>12</td>
</tr>
</tbody>
</table>

Although extracting the trigrams made it possible to compare the two corpora in
terms of frequency, the trigrams in Table 1.1 may not always make the unique features of Japanese learners’ writing obvious. Aarts and Granger (1998) apply tagging to a learner corpus and highlight the underlying features of learner writing. Learners of a second language, particularly in a classroom setting, are likely to experience a heightened awareness of grammar. Two reasons can be suggested: one is that systematic instruction of grammar is in some way involved in establishing teaching syllabuses and materials; the other is that grammar inevitably plays an essential role for learners as they generate a sentence. Tagging a learner corpus could shed light on the norms of learner language which originally derive from the grammatical instructions in the classroom, while tagging a native-speakers’ corpus aims to reveal the norms which underlie the conventional use of language. By annotating the corpora and extracting trigrams, it is hoped that this study will enable a quantitative comparison of the given corpora to be made and thus offer an insight about the peculiarities of the second language in the hands of Japanese learners.

The following section will consider the background of taggers as well as the tagging of learner corpora. The methodology of this research and the descriptions of taggers will be discussed in the third section. The fourth section presents a case study of learner corpus tagging and discussion of the results will follow in the closing section.

1. Background

Corpus annotation can provide a corpus with additional linguistic information in order to enrich the research. However, some claim that corpora should be ‘raw’ and ‘pure’, adding nothing artificial to the language, since such annotation may distort the features of a language in favour of existing linguistic theories and methods (Sinclair, 1991, Hunston 2002). Yet annotating corpora could indeed be beneficial, if ‘reliable’ and ‘clearly-definable’ information is appended to the texts (Leech, 1997). Moreover, it is maintained that tagged corpora could contribute to discovering the distinctive features of learner grammar (Aart and Granger, 1998).

Atwell and Elliot (1987) use tagging to detect errors in a text by finding that the sequence of uncommon tag pairs was caused by learner errors. They evaluate how misspelled words affect the results of tagging and develop the way in which
algorithms can generate a cohort of words similar to a mistyped word. Aarts and Granger (1998) extract the trigrams which are generated from the tagged learner corpora by Dutch, Finish and French learners. The three-word units are compared with those of native English-speaking students and a comparison of the trigrams illustrates the divergence of the interlanguage grammar from the native speaker norm.

Computer annotation began with the word-class tagging of the Brown Corpus in the early 1970s and the tagging project of the LOB Corpus launched between 1979 and 1982, aimed at achieving high rates of tagging accuracy. The bulk of early tagging software employed probabilistic methods: \textit{the Penn Treebank TreeTagger} (henceforth, \textit{TreeTagger}) implements the probability of sequences where a given word is preceded by other attributes and determines the particular tag for each word (Schmid, 1994). In terms of tagging ill-formed text, it is interesting to note that \textit{TreeTagger} is still designed to annotate misspelled words correctly (Santorini, 199).

The first version of \textit{CLAWS part-of-speech tagger} (henceforth, \textit{CLAWS}) was designed for tagging the LOB Corpus on a probabilistic basis and has since evolved to cope with tokenizing contracted forms and idiomatic phrases (Garside, 1987). Clearly, taggers inevitably reflect the linguistic theories which determine the methods of tagging. The users of taggers are thus expected to implement the methods and theories which underlie them.

2. Methodology

2.1. Data and research method

Two essay corpora were investigated, for purposes of comparison: the Showa Women's University learner corpus, about 85,000 words, which is composed of 175 essays by Japanese learners and the corpus of British A-level student essays, which is extracted from the \textit{LOCNESS} (Louvain Corpus of Native English) corpus, the size of which is approximately 80,000 words; this has been adapted to correspond with the size of the learner corpus. These corpora are designed according to the corpus design criteria of the \textit{ICLE} (the International Corpus of Learner English), which controls many variables of the corpora to allow mutual comparison.

The two corpora are tagged by \textit{TreeTagger} and \textit{CLAWS} and the trigrams of tags
are enumerated and sorted by the Perl Program in order of frequency. The overused and underused trigrams in the learner corpus, when compared to the native corpus, will be examined by reading the concordance lines in order to discover typical choices of words in the trigrams.

2.2. Tokenization

Tokenization, which assigns a token to a word unit, should be completed before the tagging of annotations, since without tokenization it might be impossible to assign tags appropriately to word units. The languages which have one-character-size spaces

1 The Perl programme was employed to extract trigrams as follows:
<Perl programme for extracting trigrams from the results of the Treetagger>
$firstline=<; $secondline=<;
($token, $firsttag, $lemma)=split(/\t/, $firstline, 3);
($token, $secondtag, $lemma)=split(/\t/, $secondline, 3);
while(<>){
    chomp $_;
    $line=$_;
    ($token, $thirdtag, $lemma)=split(/\t/, $line, 3);
    print "$firsttag $secondtag $thirdtag\n"
    $firsttag=$secondtag;
    $secondtag=$thirdtag;
}

2 Overuse and underuse are determined by comparing the raw frequencies of trigrams, that is, no statistical comparison with probability is undertaken.
between words—such as English, Spanish and other European languages—are mainly assigned one token per word, but some contend that ‘contracted forms’ may cause deviation from the one-to-one relation between a word and a token (Leech, 1997, Garside, 1987). The contraction of words primarily occurs with such verbs as be, have and modal verbs, followed by ’s, ’d or another abridged ending: am into ’m; are into ’re; have into ’ve. In addition, there are other contracted forms affecting tokenization: the enclitic form of the negative particle not as n’t, attached to the ends of verbs or modal verbs. In order to tag a text containing such contracted forms, the contracted part of the word has to be segregated from the verb as a way of inserting a one-character-size empty space directly before the contracted form: for instance, I’m into I ’m; isn’t into is_n’t.

The description of CLAWS tokenization is not available, so that the tokenization needs to be assessed by analysing the list of contracted forms (Leech and Smith, 2000). The contraction attaching to be, have, modal verbs and n’t is seemingly the identical classification for tokenization to that for TreeTagger. CLAWS can in addition assign a token to the contracted forms which are composed without the apostrophised n’t, and also the contraction which is derived from the omission of a syllable in everyday pronunciation (e.g. do you → d’you).

2.3. Tagsets

Tagsets represent the list of tags employed for tagging, where the content of the tagsets reflects the method of anatomising a language. There are various types of

3 Tokenization was carried out by the following script:

# do tokenization
$TOKENIZER +1 +s +l $ABBR_LIST |
# separate clitics from preceding words
sed -e "s/’s”’s’/ /a/g" |
- e "s/s’”s’/ /g" |
- e "s/n’t”’n’/ /t/g” |
- e "s/’re”’r’e/g” |
- e "s/’ve”’v’e/g” |
- e "s/’d”’d’/ /d/g” |
- e "s/’m”’m’/ /m/g” |
- e "s/’em”’e’m’/ /em/g” |
- e "s/’ll”’l’e”’l’/ /ll/g” |
- e ‘/”$/d’ |
tr ‘ ‘ ‘n’ |
Tagging a Japanese Learner Corpus of English and Comparing Trigrams with Those in a Corpus of British Students’ Essays

automatic and manual tag: semantic tags, syntactic tags, prosodic tags and error tags. This research seeks to describe the grammar features of Japanese learners’ writings, resulting in a concentration on the part-of-speech tagsets of TreeTagger and CLAWS (Santorini, 1991, UCREL, n.d.).

The structure of a tag is closely related to the linguistic theories and methods which are adopted when establishing the tags (Hunston, 2002). The methods of encoding tags vary in their ways of categorising annotated words. For example, prepositions are classified in the TreeTagger tagset into ‘tags for prepositions and subordinating conjunctions’, while the CLAWS C7 tagset has four subcategories for prepositions⁴.

Tags embody the classification of tagging by being represented in an acronym composed of two to five characters. In general, the formation of a tag indicates a category hierarchy by beginning with the first one or two characters which correspond to the part-of-speech of a given token: N=noun, V=verb, JJ=adjective, R or RB=adverb, DT=determiner. Note that the grammar categories are arbitrarily determined by the choice of grammar theory, so that a token may have distinct symbols or categories for a word: for example, IN=preposition in TreeTagger; II=general preposition, IF=for, IO=of, IW=with, without in CLAWS. Beneath the core category of part-of-speech is a subordinate category which attributes a ‘context-dependent’ meaning to tag symbols (Leech, 1997: 27), such that N (noun) is followed by P, which refers to ‘proper’ or double P (PP) meaning ‘personal pronoun’.

2.3.1 The TreeTagger tagset
This tagset consists of 36 tags, fewer than the CLAWS tagset. However, the size of a tagset is not necessarily important, since increases and decreases of size depend on what is emphasised for the purpose of the tagging (Leech, 1997). The TreeTagger tagset seems simple and recognisable when decoding the tags; however, the limitation of the tag variety, in terms of discouraging ambiguity, may create an overlap in categories which should be rigorously distinctive, for example, IN for prepositions and subordinating conjunctions.

---

⁴ A free CLAWS tagging service is available on the Internet (http://www.comp.lanc.ac.uk/ucrel/claws/trial.html), offering the choice of UCREL C5 or C7 tagset. The C7 tagset is well explained in an article by Leech (1997).
2.3.2 CLAWS tagset (*UCREL CLAWS7 tagset*)

The *CLAWS7* tagset consists of 141 tags, which mostly represent the attributes of tokens with three-character symbols. Some of these symbols are followed by the number which stands for singular or plural—NN1 for singular common noun, AT1 for singular article such as *a, an*. These tagsets seem to retain high analysability; however it is difficult to decode them without consulting the list of tagsets. It needs to be determined in the trade-off whether conciseness and perspicuity are worth the sacrifice of high analysability.

### Table 3  Extract of *UCREL CLAWS7* tagset (*UCREL* n. d.)

<table>
<thead>
<tr>
<th>AT</th>
<th>Article (e.g. <em>the, no</em>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1</td>
<td>Singular article (e.g. <em>a, an, every</em>)</td>
</tr>
<tr>
<td>IF</td>
<td>For (as preposition)</td>
</tr>
<tr>
<td>II</td>
<td>General preposition</td>
</tr>
<tr>
<td>IO</td>
<td>Of (as preposition)</td>
</tr>
<tr>
<td>IW</td>
<td>With, without (as preposition)</td>
</tr>
<tr>
<td>JJ</td>
<td>General adjective</td>
</tr>
<tr>
<td>NN1</td>
<td>Singular common noun (e.g. <em>book, girl</em>)</td>
</tr>
<tr>
<td>NN2</td>
<td>Plural common noun (e.g. <em>books, girls</em>)</td>
</tr>
<tr>
<td>TO</td>
<td>Infinitive marker (<em>to</em>)</td>
</tr>
<tr>
<td>VVI</td>
<td>Infinitive (e.g. <em>to give..., It will work...</em>)</td>
</tr>
<tr>
<td>VVN</td>
<td>Past particle of lexical verb (e.g. <em>bound in be bound to</em>)</td>
</tr>
<tr>
<td>.</td>
<td>Full stop</td>
</tr>
</tbody>
</table>
In this section, the methodologies, tokenization and tagsets were discussed. Before interpreting the results of tagging, it may be necessary to ascertain how the tokenizer functions and what tags are assigned. The tokenization of taggers which are defined by identical classification of contracted forms may not affect the results, when the tagging is compared. However, note that the assignment of tags differs in the two tagsets, which may closely relate to the interpretation of the tagged outcome. On this basis, the tagging of a learner corpus and an application of tagged learner corpora will be considered in the next section.

3. Case study of learner corpus tagging

3.1. Evaluation of learner corpus tagging

Practical applications of learner corpora are highly likely to involve analysing texts which may contain misspelled words. Atwell and Elliot (1987) apply the peculiarity of the ill-formed word to detecting and marking learner errors. Typing errors may result in mistakenly assigning an error word to an unmatched tag, which has a different meaning from what the learner intended. Before interpreting the tagged results, it should be confirmed whether the two taggers have assigned suitable tags to the remaining misspelled words and how they treat unidentified words in the process of tagging.

*TreeTagger* yields the results of tagging, as seen in Table 4.1, in the order of ‘token’, ‘part-of-speech tag’ and ‘lemma’. It is likely that the low flexibility of the tokenization may cause inaccurate tagging. *Cannot* is tagged as JJ (adjective) and the lemma turns out to be <unknown>, a word which may not be found in the word list with which the tagger identifies the lemma of words. Had the tokenization divided it into *can* and *not*, they could have been assigned to appropriate lemmas. Misspelled

---

5 A tagset for a learner corpus will be briefly discussed: the TOSCA-ICLE tagset is designed for tagging the ICLE corpus. Where it differs from *TreeTagger* and CLAWS is the uniformity of the tag symbols which refer to the part-of-speech and the illustration of the subcategorised part in brackets, e.g., N (sing, collect). The TOSCA-ICLE tagger is restricted in its circulation and it proved impossible to contact the authors of the tagset and an ICLE researcher. The discussion on the tagger exclusively centres on the tagset, which is available on a Kaszubski Webpage (Kaszubski, 2003).
words can be assigned a particular tag by guessing them from the surrounding words, despite showing an <unknown> lemma, which may not identify with any of the words in the lemma list.

<table>
<thead>
<tr>
<th>Table 4 Results of the Penn Treebank TreeTagger tagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>cannot JJ &lt;unknown&gt;</td>
</tr>
<tr>
<td>pm. [sic.] NN &lt;unknown&gt;</td>
</tr>
<tr>
<td>you [sic.] JJ &lt;unknown&gt;</td>
</tr>
<tr>
<td>firends [sic.] NNS &lt;unknown&gt;</td>
</tr>
<tr>
<td>not RB not</td>
</tr>
<tr>
<td>think VB think</td>
</tr>
<tr>
<td>that DT that</td>
</tr>
<tr>
<td>I PP I</td>
</tr>
<tr>
<td>though [sic.] RB &lt;unknown&gt;</td>
</tr>
<tr>
<td>that IN that</td>
</tr>
</tbody>
</table>

Misspelled words may cause error tagging. For example, the *that* which introduces a complement clause, as in *I do not think that most Japanese students are too easy going*, mainly falls into being tagged as a determiner (DT). However, the *that* which follows an ill-formed word *though* as in *I though that I would study hard in college*—maybe a mistake for *thought*—have a preposition tag (IN). The preceding misspelled word may affect the choice of tagging for the complement subordinator.

The C7 vertical format of the CLAWS tagging shows that the two principal results are ‘token’ and ‘tag’⁶. It should be noted that the tagger indicates the probability of tags. For example, a *that* which follows *think* as a complement subordinate is estimated as 63% of conjunction *that* (CST) and 37% of singular determiner (DD1) and results in its being tagged as CST.

---

⁶ The horizontal tagging of CLAWS indicates only the highest probability’s being tagged, since the tag format is designed for the application of the tagged text to the concordance software, not for analysing the results of the tagging.
Table 5  Results of UCREL CLAWS7 tagging

<table>
<thead>
<tr>
<th>Tagging</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000004 820 can</td>
<td>&gt; 56  VM</td>
<td></td>
</tr>
<tr>
<td>0000004 821 not</td>
<td>&lt; 56  XX</td>
<td></td>
</tr>
<tr>
<td>0000004 020 pm. [sic.]</td>
<td>93    RA</td>
<td></td>
</tr>
<tr>
<td>0000003 630 you [sic.]</td>
<td>06    [NN1/99] VV0/1</td>
<td></td>
</tr>
<tr>
<td>0000004 270 friends [sic.]</td>
<td>98    NN2</td>
<td></td>
</tr>
<tr>
<td>0000003 030 not</td>
<td>93    XX</td>
<td></td>
</tr>
<tr>
<td>0000003 040 think</td>
<td>93    VVI</td>
<td></td>
</tr>
<tr>
<td>0000003 050 that</td>
<td>97    [CST/63] DD1/37 RG%/0</td>
<td></td>
</tr>
<tr>
<td>0000003 290 I</td>
<td>93    [PPIS1/100] ZZ1%/0 MC1%/0</td>
<td></td>
</tr>
<tr>
<td>0000003 300 though [sic.]</td>
<td>93    [RR@/98] CS/2</td>
<td></td>
</tr>
<tr>
<td>0000003 310 that</td>
<td>96    [CST/100] DD1%/0</td>
<td></td>
</tr>
</tbody>
</table>

These elaborated tags, in part, appear to deal with the problems of tokenization and misspelled words which the TreeTagger raises. Cannot is tokenized into two words can and not, tagged modal auxiliary and not respectively, while mistyped pm. [sic.] and you [sic.] have inappropriate tags. The complement subordinator that has the right tag of conjunction that (CST), even following the tagging of an erroneous verb though [sic.] as a general adverb (RR).

The bulk of tags are properly assigned; however the misspelled words which are frequently found in learner writings may result in misleading tagging. Hence, despite the possibility that tagging may present some findings which underlie the text, verifying the matching of tags and perusing concordance lines is essential when interpreting the results of tagging a learner corpus, as well as tagging corpora in general. We will therefore examine the trigrams of tags first and then scrutinise the concordance lines.

3.2. Trigrams by TreeTagger

This section will investigate the results of trigrams of the two corpora being tagged by TreeTagger and continue the interpretation of the overuse and underuse trigrams compared with those of native writings.
Table 6  The ten most frequent trigrams and underuse of the two corpora by TreeTagger

<table>
<thead>
<tr>
<th>JP</th>
<th>Trigram</th>
<th>NS</th>
<th>Trigram</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq.</td>
<td></td>
<td>Freq.</td>
<td></td>
</tr>
<tr>
<td>1283</td>
<td>IN DT NN</td>
<td>1833</td>
<td>IN DT NN</td>
</tr>
<tr>
<td>1017</td>
<td>DT NN IN</td>
<td>1678</td>
<td>DT NN IN</td>
</tr>
<tr>
<td>581</td>
<td>NN IN NN</td>
<td>1148</td>
<td>NN IN DT</td>
</tr>
<tr>
<td>538</td>
<td>JJ NN IN</td>
<td>780</td>
<td>IN DT JJ</td>
</tr>
<tr>
<td>526</td>
<td>IN PP VBP</td>
<td>747</td>
<td>JJ IN NN</td>
</tr>
<tr>
<td>508</td>
<td>IN DT JJ</td>
<td>657</td>
<td>NN IN NN</td>
</tr>
<tr>
<td>505</td>
<td>NN SENT IN</td>
<td>426</td>
<td>NNS IN DT</td>
</tr>
<tr>
<td>504</td>
<td>NN IN DT</td>
<td>391</td>
<td>VBN IN DT</td>
</tr>
<tr>
<td>404</td>
<td>IN NN SENT</td>
<td>388</td>
<td>IN DT NNS</td>
</tr>
<tr>
<td>392</td>
<td>IN JJ NN</td>
<td>357</td>
<td>IN JJ NNS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>125</td>
<td>VBN IN DT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>208</td>
<td>IN NN SENT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>144</td>
<td>IN PP VBP</td>
</tr>
</tbody>
</table>

a) $IN+PP+VBP$ (Preposition or Subordinating conjunction + Personal pronoun + Verb, non-3rd person singular present)

This trigram shows the distinctive frequencies being considerably overused. The IN tag refers to prepositions ($in$, $on$, $at$ etc.) and subordinating conjunctions ($as$, $if$, $because$ and other subordinators) of the latter, which do not correspond to the target preposition $in$. By reading the concordance lines extracted according to the trigram, many of the IN tags turn out to be tagged not on preposition $in$, but on subordinating conjunctions, such as $although$, $because$, $if$, $that$ and the like. The tags PP and VBP, indicating personal pronouns and verbs, illustrate this set of tags as beginners of clauses. Despite the attempt to tag a phrase containing the preposition $in$, this trigram only managed to reveal the overuse of phrases which were composed of subordinating conjunction, personal pronoun and singular present verb. Yet the concordance lines extracted on the basis of the trigram show an interesting phraseology by Japanese learners: of 526 instances of $IN+PP+VBP$, the clauses starting with $if$ (168) occur more frequently than those with $that$ (152). This may imply that Japanese learners are inclined to use $if$-clauses in their essay writings. It should, however, be noted that essay
topics selected by the learners might relate to the surge of frequency of if; that is, a hypothetical question as essay topic is likely to encourage learners to argue by using if-clauses excessively. Reading concordance lines can avoid matching with different tags from what one expects, and lead to findings which may be useful and suggestive.

b) IN+NN+SENT (Preposition or subordinating conjunction + Noun, singular or mass + Sentence delimiter)
The tokenization of TreeTagger nominates a tag ‘SENT’ for full stops, while commas are assigned to ‘,’ only. This trigram seems to illustrate that the Japanese learners tend to use prepositional phrases at the ending of a sentence. The 63 instances of in phrases do not dramatically vary in the types of noun: college, future, school, society, summer, university, workplace, world predominate in the phrases. Japanese learners tend to express the time when they were studying in a college, school and university by using the phrase ‘in + INSTITUTION’. This prepositional phrase does not contain article the between the preposition and college, school, university, which implies a lack of articles in the writings by Japanese learners. Different terms for institutions of learning—college, school, university—are employed for referring to either time or place ambiguously, as shown below:

1 I will get along in the world while I would be in university. So I will stud
2 lusion, It is very comfortable for me to be in university. There are a lot
3 d to this. Seventhly why do I learn English in University. I just learn Engl
4 nglish in University. I just learn English in University. But I need to
5 the best of time. We can learn special field in university. I learn English
6 be in university. There are a lot of friends in university. And I can have a
7 ther country’s language. I studying German in university. German is difficu
8 nd I want to find the job after graduation in university. Finally, in the

By contrast, the nouns in the prepositional phrases in the native writings show considerable variety, unlike those in the learner corpus. They are not only words which refer to time and place, but also to words which signify the state which the agent of a clause goes into—advance, charge, failure, general, luxury, unemployment, which do not occur with article the. As a result, prepositional phrases which denote a metaphorical state, such as in advance, in failure, in luxury, may be difficult for learners to learn to produce in the writings of the second language.
1. PP want_VBP to_TO travel_VB in_IN advance_NN ._SENT Also
to_TO keep_VB farmers_NNS in_IN business_NN ._SENT The
dT human_NN is_VBZ still_RB in_IN charge_NN ._SENT The_D
4. N to_TO its_PP$ treatment_NN in_IN court_NN ._SENT Our_PP
to_TO these_DT changes_NNS in_IN demography_NN ._SENT
6. an_MD do_VB is_VBZ flee_VB in_IN desparation_NN ._SENT It
t of_IN course_NN ended_VBN in_IN failure_NN ._SENT There_
8. G beef_NN ,_, or_CC meat_NN in_IN general_NN ._SENT Meat
en_NNS alike_RB travel_VBP in_IN luxury_NN ._SENT British
10. _ CC more_JJR importantly_RB in_IN manufacturing_NN ._SEN
t single_JJ market_NN is_VBZ in_IN operation_NN ._SENT It_P
VBD instantly_RB divided_VBN in_IN opinion_NN ._SENT The_
13. B down_RP ;_, resulting_VBG in_IN unemployment_NN ._SEN
14. JJ yard_NN is_VBZ already_RB in_IN use_NN ._SENT The_DT

c) VBN+IN+DT (Verb, past participle + Preposition or subordinating conjunction + Determiner)
This trigram is underused by Japanese learners; that is, trigrams of this kind in the learner corpus are much more rarely than in the native corpus. The trigrams by the learners interestingly show a regular patterning of multi-word units together with particular past participles: be interested/located/spoken/written in. The variations of the past participles in the trigram turn out to be smaller that those which occur in the British students’ corpus, which may result in the underuse of this trigram. Of the limited variety of the sequence of past participle + in + determiner, the dominant frequency of interested in may result from the materials which learners use in classroom: a textbook for Japanese learners of English presents be interested in as a likely set phrase (Kasashima et al. 2006). Such input provided in the classroom is likely to affect the outputs of learners; in some cases learners’ outputs may be formulated by the instruction to use a pseudo-set phrase.

VBN+IN+DT in the learner corpus
1. I_PP am_VBP interested_VBN in_IN another_DT country_NN cul
2. T I_PP ‘m_VBP interested_VBN in_IN another_DT country_NN ._
3. e_PP can_MD interested_VBN in_IN another_DT culture_NN and
d_CC are_VBP interested_VBN in_IN the_DT final_JJ information
4. I_PP am_VBP interested_VBN in_IN the_DT first_JJ type_NN ._
5. ENT It_PP ‘s_VBZ located_VBN in_IN the_DT central_JJ Bali_NP
Tagging a Japanese Learner Corpus of English and Comparing Trigrams with Those in a Corpus of British Students’ Essays

The components of this trigram among British students appear to be chosen differently from those among Japanese learners. The use of the past participles in the writings by the British seems to fall into two subcategories: where the verbs recurrently follow an auxiliary verb + be and where they can follow either be or a noun. Adopted, involved, left, used and other past participles which frequently occur in the British student corpus could be introduced to learners as frequent words in passive clauses by native speakers. Such suggestions may enrich learners’ expressions in their outputs.

**be VBN in DT in the native corpus**

```
be VBN in DT in the native corpus
```

**be VBN in DT or NN VBN in DT in the native corpus**

```
be VBN in DT or NN VBN in DT in the native corpus
```
The comparison of learner and native examples of the trigrams generated from the tagged words by TreeTagger illustrates that the patterning of multi-word units may be both overused and underused. However, the tag trigram a) has not managed to show three-word sequences containing preposition in. The following section will investigate the way in which the trigrams are tagged by CLAWS.

3.3. Trigrams by CLAWS

CLAWS tags the two identical corpora being tagged by TreeTagger in the previous section. The CLAWS tagger, which is made up of more tags than TreeTagger, may shed light on features which were not previously elicited by the TreeTagger tagging.

Table 7 The ten most frequent trigrams and underuse of the two corpora according to CLAWS

<table>
<thead>
<tr>
<th></th>
<th>JP Freq.</th>
<th>Trigram</th>
<th></th>
<th>NS Freq.</th>
<th>Trigram</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>518</td>
<td>II AT NN1</td>
<td>763</td>
<td>II AT NN1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>283</td>
<td>II NN1 .</td>
<td>361</td>
<td>NN1 II AT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>205</td>
<td>TO VVI II</td>
<td>266</td>
<td>JJ NN1 II</td>
<td></td>
</tr>
<tr>
<td></td>
<td>196</td>
<td>VVI NN1 II</td>
<td>261</td>
<td>II AT JJ</td>
<td></td>
</tr>
<tr>
<td></td>
<td>196</td>
<td>NN1 II AT1</td>
<td>235</td>
<td>II AT1 NN1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>191</td>
<td>II JJ NN1</td>
<td>207</td>
<td>VVN II AT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>183</td>
<td>NN1 II NN1</td>
<td>175</td>
<td>AT NN1 II</td>
<td></td>
</tr>
<tr>
<td></td>
<td>174</td>
<td>JJ NN1 II</td>
<td>170</td>
<td>II AT1 JJ</td>
<td></td>
</tr>
<tr>
<td></td>
<td>170</td>
<td>NN1 . II</td>
<td>168</td>
<td>II JJ NN2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>161</td>
<td>NN1 II AT</td>
<td>148</td>
<td>NN2 II AT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>VVN II AT</td>
<td>102</td>
<td>TO VVI II</td>
<td></td>
</tr>
<tr>
<td></td>
<td>91</td>
<td>II NN1 .</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>VVI NN1 II</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
d) **II+NN1+. (General preposition + Singular common noun + Full stop)**

As mentioned in the section on tagsets, these two taggers each have original tags which do not correspond in the system of coding. In decoding and interpreting the results of tagging, the unlikeness of tag representation should be borne in mind. However, this trigram does represent the same three-word units as are encoded as IN+NN+SENT in *TreeTagger*. Although the representations are coded by different symbols, the words of the prepositional phrases are individually identified in both tagged trigrams. Interestingly, this trigram occupies the second place for frequency—113 instances of *in* out of 283 preposition tokens, which is compared to ninth place in the *TreeTagger* trigram results—62 instances of *in* out of 404 instances. This divergence illustrates that the specification of a tag can affect the interpretations of tagging. Therefore, a close examination of tagsets may be essential before interpreting the outputs by any particular tagger.

e) **TO+VVI+II (Infinitive marker [to] + Infinitive + General preposition)**

This trigram does not appear to be on the frequency list of *TreeTagger*, although the divergence of the tagsets does not affect the chance of finding this trigram. It is followed by nouns which illustrate typical features in learner writings, seemingly related to the choice of verbs. The nouns, preceded by *in*, refer to a sequence with *get, live, spread, study, survive* and *work*. These verbs preceding *in* suggest that the preference of learners is to represent the locative meaning of the preposition. At the same time, *in English* often recurs after *communicate, say* and *speak* which closely relate to the utterance of language. As a result, Japanese learners are likely to grasp the meanings of *in* in terms of location and language, while the other meanings might be too unfamiliar for the learners to use in their writings.

to VVI in place

1. NN2 ._. Hurryup_VV0 to_TO get_VVI in_I1 this_DD1 boat_NN1 !_! "_" But_  
2. enger_NN1 tried_VVD to_TO get_VVI in_I1 fullboats_NN2 ,_, Rowe_NP1 m  
3. e_VBI difficult_JJ to_TO live_VVI in_I1 this_DD1 world_NN1 ._. So_RR  
4. 1 always_RR good_JJ to_TO live_VVI in_I1 a_AT1 hospital_NN1 with_IW m  
5. We_PPIS2 have_VH0 to_TO live_VVI in_I1 these_DD2 situation_NN1 ._. To  
6. AT society_NN1 ._. To_TO live_VVI in_I1 a_AT1 nation_NN1 society_NN1  
7. nt_JJ for_IF us_PPIO2 to_TO live_VVI in_I1 the_AT twenty-first_MD century  
8. PH1 is_VBZ apt_JJ to_TO spread_VVI in_I1 world_NN1 ._. Many_DA2 foreign
to VVI in English

f) VVI+NN1+II (Infinitive + Singular common noun + General preposition)

There are 196 instances of this trigram, of which 38 instances contain preposition in. The 31 instances out of 38 of the trigrams contain English as object, following such verbs as learn, speak, study, teach and use. It is likely that the learner may be inclined to employ a patterning to learn/speak/study/teach/use English in. In addition, this trigram appears to be followed often by future and world, together with junior high school, college and university, all institutions. These components of the prepositional phrases are identified with what has been discovered in the TreeTagger trigram of b) IN+NN+SENT (Preposition or subordinating conjunction + Noun, singular or mass + Sentence delimiter) and the CLAWS trigram of e) II+NN1+. (General preposition + Singular common noun + Full stop). It could thus imply that the learner frequently employs a trigram of VVI+NN1+II which precedes another multi-word unit: VVI+NN1+II+NN1+. (full stop). Furthermore, this unit is found to co-occur with need to, which can expand the range of this set phrasing. This
Tagging a Japanese Learner Corpus of English and Comparing Trigrams with Those in a Corpus of British Students’ Essays

expansion of a set phrase agrees with the features of the idiom principle, as stated, that a semi-fabricated phrase has an indeterminate boundary (Sinclair 1991: 111). The idiom principle focuses on the features of English used by native speakers, while this finding may endorse the application of the idiom principle to a learner language.

Table 8 Recurrent patterning of VVI+NN1+II+NN1+. in the learner corpus

<table>
<thead>
<tr>
<th>TO</th>
<th>VVI</th>
<th>NNI</th>
<th>II</th>
<th>NN1</th>
<th>Full Stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>begin</td>
<td>learn</td>
<td>English</td>
<td>future</td>
<td>school</td>
<td>.</td>
</tr>
<tr>
<td>start</td>
<td>speak</td>
<td>in</td>
<td>college</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>need</td>
<td>study</td>
<td></td>
<td>university</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>want</td>
<td>use</td>
<td></td>
<td>world</td>
<td>.</td>
<td></td>
</tr>
</tbody>
</table>

VVI+NN1+II
1 " I_PPIS1 will_VM learn_VVI English_NN1 in_II junior_JJ high_JJ school_NN1
2 start_VV0 to_TO learn_VVI English_NN1 in_II junior_JJ high_JJ school_NN1 .
3 do_VD0 I_PPIS1 learn_VVI English_NN1 in_II University_NN1 _. I_P
4 desire_VV0 to_TO speak_VVI English_NN1 in_II workplace_NN1 ,_, we_PPIS2 s
5 need_VV0 to_TO speak_VVI English_NN1 in_II company_NN1 in_II Japan_NP1
6 ease_VV0 to_TO speak_VVI English_NN1 in_II the_AT world_NN1 ._. Many_
7 need_VV0 to_TO speak_VVI English_NN1 in_II the_AT future_NN1 ._. So_
8 ided_VVD to_TO study_VVI English_NN1 in_II collage_NN1 ._. Other_JJ rea
9 egan_VVD to_TO study_VVI English_NN1 in_II junior_JJ high_JJ school_NN1
10 have_VH0 to_TO study_VVI English_NN1 in_II junior_JJ high_JJ school_NN1
11 start_VV0 to_TO study_VVI English_NN1 in_II junior_JJ high_JJ school_NN1
12 nted_VVD to_TO study_VVI English_NN1 in_II this_DD1 university_NN1 ._.
13 want_VV0 to_TO teach_VVI English_NN1 in_II the_AT future_NN1 ._. If_CS
14 us_PPIO2 to_TO use_VVI English_NN1 in_II the_AT future_NN1 ._. We_PPIS

g) VVN+II+AT (Past participle of lexical verb + General preposition + Article)
An underused trigram in CLAWS to be compared to the native usage is VVN+II+AT, which represents the equivalent three-word unit to VBN+IN+DT (Verb, past participle + Preposition or subordinating conjunction + Determiner) in the TreeTagger tagging. However, the extracts of this unit from the native corpus do not show typical patterning of phrases, such as involved in the, which has been found in the
results of *TreeTagger*. This is because *TreeTagger* tags *involved* as a past participle verb whereas *CLAWS* assigns the word to a JJ (adjective) tag, resulting in its not being identified as a past participle verb. This discrepancy may endorse the view that methods of assigning words to tags vary between taggers and the characteristics of a tagger may lead to one system’s using a different tag from another.

**VVN+II+AT in the native corpus**

```
1  CST could_VM be_VBI adopted_VVN in_II the_AT U.K._NP1 It_PPH1 would
2  NN1 should_VM be_VBI banned_VVN in_II the_AT United_NP1 Kingdom_N
3 _AT brain_NN1 is_VBZ encased_VVN in_II the_AT skull_NN1 ,_, but_CCB
4  uld_VM become_VVI engulfed_VVN in_II the_AT new_JJ Single_JJ Europe
5  mb_NN1 ._. When_CS formed_VVN in_II the_AT womb_NN1 they_PPHS2
6  ind_NN1 can_VM be_VBI banned_VVN in_II the_AT United_NP1 _AT brain_NN1
7  d_VVN to_TO be_VBI implanted_VVN in_II the_AT mother_NN1 years_NNT
8  not_XX originally_RR included_VVN in_II the_AT treaty_NN1 )_) ,_, gives
9  N1 ,_, originally_RR invested_VVN in_II the_AT Crown_NN1 ,_, today_R
10  PH1 may_VM be_VBI irradiated_VVN in_II the_AT cooking_NN1 process_N
11  ry_RG much_RR been_VBN left_VVN in_II the_AT dark_NN1 as_CSA their_
12  AT1 fuss_NN1 was_VBDZ made_VVN in_II the_AT newspapers_NN2 over_II
```

This section shows that the two taggers generate seven types of trigram to extract typical multi-word units. It is also clear that the two taggers retain their unique attributes, as should not be forgotten in the process of interpreting the trigrams.

4. Discussion and conclusion

This study had the purpose of making a quantitative and qualitative comparison between a Japanese learner corpus and one by British students. Annotating the learner corpus has enabled us to compare the environment of *IN* and has shown the overuse and underuse of trigrams containing the preceding and following words of a certain preposition. As a result of this, there appear to be two distinctive features in the writings by Japanese students: the overuse of the prepositional phrase containing *IN* in the final position of a sentence; the overuse of *to*-infinitive before *IN*; and the underuse of past participle before *IN* as a prepositional phrase. Reading the concordance lines of the given trigrams reveals that Japanese learners tend to use *IN*
before the nouns denoting places and times, such as college, school, society, university, workplace, world, future and summer (Trigrams b and d). Likewise, the verbs representing the actions in a location frequently occur with to-infinitive and IN (Trigram e), while at the same time the verbs signifying speech activity—say, speak, communicate—recur with IN in Japanese learners’ writings (Trigram e and f). Note that the topics which are chosen by learners for the compilation of a learner corpus are highly likely to correlate to the frequency of certain words.

Tagging the Japanese learner corpus in comparison to the British students’ corpus has shed light on familiar and unfamiliar sequences co-occurring with IN preposition to Japanese learners of English. A close reading of the concordance lines of Trigram d) and f) has led to a recurrent patterning by Japanese learners: for instance, . . . need to learn English in future. This finding endorses the ‘idiom principle’, which is based on the analysis of native speakers’ text: ‘[m]any phrases have indeterminate extent’ (Sinclair, 1991: 111–2). In this respect, Japanese learners show the same features of phraseology as native speakers do. By contrast, the underuse of the trigrams containing a past participle has appeared likely to be due to the restricted vocabularies of -ed in a passive clause. This limited variation conforms insufficiently, in comparison to the native speakers’ corpus, to another phenomenon of the idiom principle: ‘[m]any phrases allow internal lexical variation’ (ibid. 111–2). More descriptions on the phraseology of learner language are required in order to account for the creativity and strategies of learners.

As has been seen, annotating a learner corpus may enrich the comparison with a native-speaker corpus from a quantitative viewpoint and may reveal a tendency of language, which a ‘raw corpus’ (Leech, 1997: 4) that is, a non-annotated corpus, fails to detect. Acknowledging the benefits, it should be remembered that a tagger is chosen according to the target of research. The IN tag of TreeTagger may not be suitable for examining preposition in phrases (Trigram a). Thus, a careful examination of tagsets and concordance lines may avoid making misleading interpretations of annotation. It should be noted that misspelled words are likely to affect the tag matching; however, the drawback may be to some extent overcome by scrutinizing the tagged concordance lines. Admittedly, this limitation needs investigating and addressing in a further study.

Differences between the written production by Japanese learners and by British native speakers have been examined in this study. Drawing attention to such
differences may encourage language instructors in classroom to raise learners' awareness of the internal lexical variation in a phrase and present learners with more variations to promote greater flexibility. In addition, contrasting the phraseology of a learner group may help teachers to know which are the common features influenced by the shared settings of the group, such as materials, teaching method and the transfer from the first language. As regards learners, in particular at an advanced level, the comparison of their phraseology with native ones could help an autonomous study of language learning by eliciting underlying features of their output.

References


Tagging a Japanese Learner Corpus of English and Comparing Trigrams with Those in a Corpus of British Students’ Essays