Part of speech tagged asynchronous CMC: Comparing native, non-native, and newspaper English

Craig Dennis Howard
Indiana University Bloomington
cdh@indiana.edu

Developing learner proficiency in L2 CMC presents unique problems not addressed in traditional course materials. The pedagogical value of learner corpora is well documented; however, corpora of learner CMC are less commonly investigated. Tagging of parts of speech and comparing L2 and native speakers’ samples shed light on the differences between written and computer-mediated language when compared with samples of more traditional texts. In this study, the author used a trained version of a Brill part of speech tagger to tag, study and analyze three corpora’s trigrams. The three corpora were: (1) asynchronous native speaker CMC, (2) non-native speaker CMC, and (3) a set of newspaper articles. The analysis found that the non-native speakers shared the newspapers’ characteristics of repeated trigrams, but dissimilar part of speech trigrams. It also found that the non-native speakers’ CMC shared a number of popular part of speech trigrams with the native speaker CMC corpus, hinting at a structural difference in computer-mediated discourse. The article questions if commonly used pedagogically texts, such as newspaper articles, can address the uniqueness of CMC and are as “authentic” as we generally consider them to be. Implications for teaching and L2 materials development are discussed.

1. Introduction

A number of scholarly journals outside computational linguistics have run special issues to investigate the pedagogical applications of using corpora for learning and learning designs (Tribble & Barlow 2001). The most common uses of corpora are to aid instruction with native speaker examples and provide collocations to inquisitive learners; however, the greatest demands in the field of Computer-assisted language learning (CALL) are not for lexical collocations, but for computers to provide automatic native speaking interlocutors and error correction for students (Atwell et al. 2000). Thus computational linguistics’ greatest potential contribution to L2 learning is not providing examples of target usage, but rich descriptions which aid in our understanding of native speakers’ online discourse rules of practice. Teachers and learners will benefit most from
rules and descriptions, not examples. In other words, the language learning exercises of the future will not come from looking at examples, but reacting in simulated discourse. This is important because the venue where learners will use their skills will be online (Warschauer 1999, Wesch 2007).

Computer-mediated communication (CMC) complicates the design of language learning curricula and materials with its alternative lexis and various genres (Baron 1998, 2000, Cherny 1999, Herring 2008). Focused curricula to address these genres are needed to teach learners how to communicate in these modes. Rich descriptions of target proficiencies and learner obstacles can identify areas which need to be taught. L2 learners must learn to recognize conventions of each mode to communicate well. Conventions with regard to emoticons, alternative punctuations and alternative spellings cannot be assumed to be simply transferred when coursework targets only tradition writing or spoken proficiencies. Furthermore, mediated language is less of a static phenomenon than a dynamic one; we need techniques and approaches to describe it, replicate it, and teach it in order to bring learners into the fold.

There have been a number of attempts to classify the exact nature of virtual communication. Baron (1998) argues that email is a kind of written pidgin, others have called it “interactive written discourse” (Ferrara et al. 1991), or even a kind of diseased language (Thurlow 2006). Independent of the various ways we may choose to identify and classify the written texts of the Internet, each CMC mode will continue to have its own unique characteristics. Herring (2001) argues that each modality of computer-mediated language contains non-standard features, yet few of these are due to ignorance of standard language forms. Herring (2007) breaks the various CMC modalities down into a faceted classification using combinations of mode and situation. Modes are defined by the technological affordances, such as synchronous versus asynchronous and persistent versus irretrievable text, whereas situations are defined by the interlocutors themselves, by their age, gender, social position in the group and so on. Only by classifying CMC studies by both facets can we expect to come to conclusions about how the modes and situations impact discourse online (Herring 2007). With the average student writing more than 12 times as much text for email as they do for school (Wesch 2007), the demand for proficiency in CMC will surely increase for non-native speakers. A methodical investigation of asynchronous CMC should reveal characteristics of discourse to inform not only our materials, but the tools we use to make those materials, whether dynamic or static.
One approach to creating a richer description of the characteristics of
discourse between different modes of communication is comparisons of
part of speech sequences. Part of speech (POS) taggers are simple scripts
which tag each item, usually words, with a part of speech and then
calculate the frequencies of sequences of these parts of speech. A
sequence of three parts of speech is called a trigram. The frequencies of
trigrams within a text describe a grammatical-structural characteristic of
that text (McEnery, Xiao & Tono 2005).

Using rich descriptions of the grammatical structure of discourse types,
educational materials and digital scaffolds could be developed. Therefore,
I ask the following questions of in an investigation of online discourse:

1. Can a POS tagger reveal insights into differences between
native and nonnative speaker asynchronous CMC?

2. In terms of part of speech trigrams, how distant are
traditional materials, such as newspaper articles, from CMC
norms? To what extent do CMC trigrams, both from native and
non-native speakers, mirror the trigrams found in a regular corpus
of written international newspaper English?

It was hypothesized that a POS trigram investigation of three sample
corpora would reveal grammatical-structural differences between texts. It
was assumed that investigating specific modes of CMC might allow a
more thorough description of new CMC grammars through the
examination of part of speech patterns. Richer descriptions of non-native
CMC are expected to inform theories of CMC behavior, and result in
practical solutions to educational challenges in the teaching of CMC
proficiencies.

2. Previous studies

A number of studies contextualize these questions. Using the frequencies
of different word classes, Heylighen and Dewaele (2002) successfully
distinguished different genres of language, identifying a formality
continuum which they labeled “F-score”. They reasoned that nouns,
adjectives, articles and prepositions are more frequent in low-context
genres of language and from this found that the frequency of these words
decreases as unambiguous communication becomes more important, and
vice versa. Written communication is low context by definition when
compared with speech because all paralinguistic information is stripped—
CMC is written (typed). This suggests that parts of speech patterns can
characterize CMC, and identifying the genre through part of speech tags is a reasonable goal. Of course, CMC registers have been identified using other means (Werry 1996). Part of speech sequences of competent users might model the CMC mode, and provide descriptors that pedagogy can apply. Tagging can also describe the overall behavior on a discussion forum. Word class categories have been used to describe interactional “presence” (Kramer et al. 2006) and spoken versus written characteristics of CMC formats (Ko 1996).

Also, automated word classifications have shown aspects of learner language and been used to make inroads into the development of L2 pedagogy. Tono (1999) presents an error taxonomy for successive levels of Japanese English language learners which he built by investigating learners’ part of speech tags compared to native speakers’ tags. McEnery (2005) mapped out learner errors according to language ability levels and created a continuum of error types and a sequence of morphemic acquisition. Using these sequences paired with frequency lists of words found in email corpora, designers of language learning curricula have focused L2 materials to better prepare language learners for communication using new media (Howard & Van Moere 2002, Howard 2004).

The contribution of learner corpora to second language learning includes studies on recognition of L1 transfer and language interference as well. Aart and Granger (1998) conducted a study using four learner corpora to develop error taxonomies of specific learner groups; Dutch, Finnish, and French learners of English were grouped into stages of development according to the part of speech tags. Tono (2000) tagged a series of Japanese learner corpora using the CLAWS (Constituent Likelihood Automatic Word-tagging) tagger and ran a similar study of the top 20 trigrams in comparison to the European learners, finding that the Japanese learners did not share similar errors. In comparison to the European students, the Japanese learners underused prepositional patterns, overused verb related trigrams, and underused modal auxiliaries and articles (Tono 2000).

Research has advocated a manipulation of the taggers according to native or non-native characteristics of the writers. Atwell (1987) looked at the CLAWS tagger and measured its precision in tagging errors written by non-native speakers. He counted the number of errors the tagger missed, the tagger’s recall, and number of errors the tagger correctly marked as errors. He found that different types of texts produced different error rates in four areas, and that non-natively written texts varied greatly from other types. Atwell found that the non-native English writers’ distribution of
errors included 4% in the non-word category (e.g., ‘errorz’), 48% in the grammatical errors category (e.g., run-ons and fragments), 12% were valid word forms with abnormal grammar (e.g., long distance number agreement), and 36% were lexically and syntactically valid forms which were actually errors but would normally require semantic analysis for detection (e.g., he had *sown* her up with a needle) (Atwell 1987). He reasoned that the CLAWS tagger could be set to different thresholds between precision and recall to more accurately tag errors in different types of texts.

Error taxonomies are not the only use of CMC corpora. In a study of the non-standard usage present in Catalan university email, CMC features were found to be responsible for bad machine translations (Climent et al., 2003). They evaluated the performance of a machine translation system and classified non-standard features that cause translation problems, quantifying the frequency and distribution of these features, and interpreting the results. They found three major areas of machine translation errors due to non-native usage: (1) unintentional non-standard features, (2) intentional non-standard features, (3) terminology specific to CMC.

POS tagging has been used to inform automated systems with regard to the nature of a given set of language. POS tags can be used to inform synthetic readers, elsewhere called text-to-voice, to vary intonation and pinpoint pauses to make audio text more understandable (Taylor & Black 2002). However, this is made possible by data which has been annotated for this purpose and no CMC corpus of descriptive annotations is currently available. These studies lend some credence to the argument that a corpus of POS tagged CMC will be needed in order to design a responsive feedback system for second language learners. A particular obstacle to creating a CMC tagged corpus which could inform such a response system is the number of different CMC modes. Classifying the nature of different modes of CMC into situational and medium factors may allow for a better analysis (Herring 2007). The problem of correct or incorrect usage can then be investigated in terms of appropriateness for the mode, rather than in terms of grammaticality. These more precise empirical methods offer an opportunity to hone our understanding of specific phenomena and apply practical solutions (Meurer 2007, Sampson 2005).

The use of taggers goes beyond distinguishing correct and incorrect language forms. In order to address the compounded complications of learner language and CMC, Heift (1998, 2003) presents a domain specific approach where no distinction between grammatical and ungrammatical syntax is made by the system; rather, the application focuses on learner
language as compared to learner level, determining if it is inside or outside of the parameters of certain learner domains. Using this approach, the learner is modeled and feedback can be focused to learner’s zone of proximal development or ZPD (Vygotsky 1978).

3. Data selection

This investigation of learners’ CMC using a POS tagger began with the collection of two corpora of approximately 8500 words each, one of competent English language users including native and proficient non-native users, and one of Thai university learners of English. The sample of proficient CMC (hereafter called the “native” sample) contained 8,464 words, and the learner (hereafter called the “non-native” sample) contained 8,730 words. As a comparable measure of non-CMC written language, a corpus of approximately the same size (10,997 words) of international English newspaper articles was collected.

The sample of asynchronous native speaker CMC was composed of forum posts to a discussion thread about teaching written by English teachers from the UK, Australia and the USA. Of course, some of these teachers may have been non-native speakers, but the discussion topic pre-supposes that they were competent enough users of the language that they would seek out a teaching discussion board where the lingua franca was English. It was assumed that their discussions would occur predominantly in internationally standard versions of English. These data were publicly available through a teaching and language learning website.¹ The mode of discussion was a forum, which used a common php bulletin board style software platform on the server side (www.phpbb.com).² Phpbb discussions are persistent and asynchronous, where posts are displayed in reverse chronological order with the most recent first, allowing a user to page back through the discussion. Subject lines can be changed without creating a new thread, and appear above each posts’ text, informing the reader if it is a response or an unsolicited post. Subject lines were not included into the corpora as they were assumed to be a uniquely different language form because they have a different grammatical structure than the written discourse investigated in this study. Subject lines are repetitive, often lack declensions of verb forms, and may even be written by the previous author of the message string, rather that the author of the content of the message.

¹ http://forums.eslcafe.com/teacher
² http://www.phpbb.com

115
The non-native speaker sample was taken from threads on the same bulletin board, where Thai students discussed, in English, topics assigned to them by a teacher. The forum was moderated by teachers of the sections which assigned the online discussions. Initial prompts were posted by teachers, but not included in the sample. The prompts were, “How can you best use the internet to learn English?” and “What do you do at school with your friends?” The Thai students were part of several sections of University courses taking place in Thailand where the forum was used as a support medium for students to converse in English outside of class. Threads other than the ones selected were also used for these tasks, but were not included in this corpus. Both the native speaker CMC corpus and the non-native (Thai students writing in English) discussions took place over a series of weeks during the summer of 2007.

International newspaper articles were randomly chosen after meeting a length criteria of at least 50 words. News blurbs, press releases, headlines, bylines and advertisements were not included. The source of the newspaper corpus was the Associated Press Worldstream English Service: Associated Press Data File which is a corpus available through the University of Pennsylvania at the Linguistic Data Consortium (http://www.ldc.upenn.edu/) and is designed to be representative of a standard English commonly employed internationally. The newspaper articles were all dated in 2007.

4. Methods

After collection, each corpus was cleaned and tokenized. All message headers, time stamps, teacher encouragements or announcements in the non-native sample, and metadata were removed before tokenization. A Brill tagger was downloaded and trained with the Susanne Corpus, which is a 130,000-word electronic sample of English annotated by hand to train taggers for identifying parts of speech (Sampson 1995). This corpus is made up of a variety of English texts, both spoken and written, and cleaned of local variants of English to provide a tagger with a lexical and grammatical sample representative of the standard forms of the language, avoiding local conventions of usage. The tagger and the training corpus are available online through the Susanne project website. After the tagger was trained, I tagged the three corpora for parts of speech. I then quantified trigrams of POS tags for each data set using a tool within the tagger. A selection of the twenty most frequently occurring trigrams were extracted, following the strategy employed by Tono in his study of

3 http://www.grsampson.net/Resources.html
Japanese learner morphemic acquisition sequences (2000). The number of unique trigrams was also calculated for each data set. Trigrams including the *period-space* digram sequence were not included in the POS calculations because they were highly common in all data sets and included only one actual word.

The newspaper corpus was slightly larger than the other two corpora. To account for this, the two other corpora’s word counts were averaged, and the resulting number was divided into the word count of the newspaper corpus. This resulted in the number .789 which was then used to normalize the raw scores to make the newspaper scores comparable to the other two. Trends in the data were then analyzed by the author.

The areas of focus were POS trigram repetitions and commonality of repeated POS trigrams between the three corpora. Repetitions of trigrams for each corpus were placed on a continuum moving towards uniqueness. Common POS trigrams between the three corpora were also identified and compared in their relationships among the three corpora.

### 4.1. Challenges

Tokenizing required spacing to be standardized. In the non-native text, the end-of-sentence periods often lacked a space before the next sentence—a non-native CMC convention which might have altered the tagger’s results. Also, spaces needed to be inserted between collected messages; otherwise the first word of a message and previous messages’ final punctuation would run together as one sequence. Excessive spacing presented another problem. Too many spaces between data items would unfairly influence the tagger and prevent accurate tagging. The author manually checked each sentence in the samples for these issues. No automatic feature could insert these spaces because URL (universal resource location, e.g. http://~) addresses would be broken up if this were done. Knowing how the URLs would be tagged was important to the study because it was assumed they would contribute to the overall calculation of trigram frequencies. URLs are, after all, virtual places, at least within the primary metaphors of the Internet, so it was hoped they would be tagged as nouns. Other punctuation included within the data was left as is, except emoticons. Emoticons in the non-native sample were so frequent and expressive that the symbol sequences needed to be added to the tagger. A separate tag was created for eight different sequences of punctuation and parentheses. All emoticons were checked that they adhered to one of the sequences the tagger would recognize. This allowed the POS tagger to address emoticons as a separate lexical item.
5. Results

I present the results of this study in two sections, first reporting the frequency of repeated trigrams, and then providing a comparison of trigrams appearing in the most frequent top twenty.

5.1. Frequency of repeated part of speech trigrams

The non-native sample contained the largest number of repeated trigrams by a ratio of approximately 7:6:4 when compared to the newspaper and native speaker top twenty most repeated trigrams, respectively. The most common non-native speaker trigram appeared 71 times, while the most common newspaper trigram appeared 56 times and the most common native speaker trigram appeared 35 times. This relationship remained constant over the course of the most frequent twenty trigrams. The newspaper results were consistently between the two other sets of scores in terms of POS trigram repetitions. The native speakers’ sample was spread over a wider number of trigrams; they used unique sequences of parts of speech more extensively than the non-native speakers. Figure 1 shows a comparison of the frequencies of the twenty most repeated POS trigrams for the native, non-native and newspaper samples.

The non-native speakers’ 20 most frequent trigrams totaled 715 repetitions as compared to the newspapers’ 578 and the native speakers’ 407. The non-native speakers, using a larger percentage of trigrams in repetition, had fewer unique trigrams overall. In both the top and lower end of the most repeated twenty trigrams, the non-native speakers approximately doubled the native speakers in frequency of repetitions of trigrams. The results for the newspaper corpus fell between the two.
5.2. Frequency of unique POS trigrams appearing less than ten times

Supporting evidence for the diversity of native speaker POS sequences is provided by the number of trigrams that appeared rarely in each sample. The samples were roughly the same size and the total number of trigrams found by the tagger was proportionate among the three corpora. However, there were 3479 unique trigrams in the non-native corpus but 5001 in the newspaper sample and 5023 in the native speaker sample. The native speaker CMC contained many more trigrams which appeared only once. The non-native speakers used a total of 2,277 trigrams only once in their sample, for 65% of their trigrams. However, the native speakers used a total of 3,756 trigrams only once, for 74%. Thus the native speakers used a larger variety of trigrams, with fewer repetitions. Figure 2 depicts frequencies of these once-occurring and uncommon trigrams. What may not be fully visible in the graph is that the lines intersect between three and four repetitions; therefore, I have included the raw numbers underneath to show the progression more clearly.

From five occurrences onward, the non-native speaker sample showed more repetitions of trigrams. The size of the two samples differed by only 266 words, so it is unlikely that a difference in sample size determined the point where non-native repetitions meet with the same frequency as native
speaker uniqueness. The non-native speaker sample repeated more and had fewer unique trigrams, while the native speaker corpus had more cases of trigrams used four times or less. The newspaper article text fell between the two CMC corpora in both repetitions and unique trigrams.

![Graph showing frequency of unique trigrams appearing 1-10 times in each corpus]

**Figure 2. The frequency of unique trigrams which appear only 1 to 10 times in each corpus**

5.3. A Comparison of POS sequences between corpora

Two trigrams appeared in all three corpora’s top sixteen. The trigram *preposition-definite article-singular noun* (e.g. *to the boy*) appeared in all three corpora’s top three, and *singular noun-coordinating conjunction-singular noun* (e.g. *boy and girl*) in all three corpora’s top 16. Four POS trigrams in the top twenty were shared between the native speakers CMC and the newspaper only. They were singular *noun-preposition-definite article, indefinite article-adjective-singular noun, definite article-adjective-singular noun, and adjective-singular noun-preposition*. The newspaper shared no POS trigrams exclusively in common with the non-native speakers, however the native CMC board did. Table 1 shows the top 20 trigrams in the three corpora.
Table 1: Most common parts of speech trigrams compared across three corpora

<table>
<thead>
<tr>
<th>Native Speaker Common Trigrams</th>
<th>Non-native Speaker Common Trigrams</th>
<th>Newspaper Common Trigrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. indefinite article, adjective, singular noun</td>
<td>prep, def article, singular noun</td>
<td>prep, definite article, singular noun</td>
</tr>
<tr>
<td>2. prepositional, possessive personal pronoun, singular noun</td>
<td>personal pronoun &quot;I&quot;, modal verb, basic form verb</td>
<td>indefinite article, singular noun, “of”</td>
</tr>
<tr>
<td>3. prep, definite article, singular noun</td>
<td>personal pronoun &quot;I&quot;, basic form verb, &quot;to&quot;</td>
<td>definite article, adjective, singular noun</td>
</tr>
<tr>
<td>4. present tense verb, “to”, present tense verb</td>
<td>adjective, singular noun, period</td>
<td>indefinite article, adjective, singular noun</td>
</tr>
<tr>
<td>5. present tense verb, definite article, singular noun</td>
<td>singular noun, prep, singular noun</td>
<td>preposition, definite article, adjective</td>
</tr>
<tr>
<td>6. personal pronoun “you”, modal verb, present tense verb</td>
<td>singular noun, coord conj, singular noun</td>
<td>singular noun, period</td>
</tr>
<tr>
<td>7. singular noun, “to”, present tense verb</td>
<td>sub conj, personal pronoun &quot;I&quot;, base form verb</td>
<td>singular noun, definite article, preposition</td>
</tr>
<tr>
<td>8. plural noun, period, present tense verb</td>
<td>emoticon, emoticon, emoticon</td>
<td>adjective, singular noun, singular noun</td>
</tr>
<tr>
<td>9. indefinite article, singular noun, prep</td>
<td>Yes/No, period, personal pronoun &quot;I&quot;</td>
<td>singular noun, “of”, definite article</td>
</tr>
<tr>
<td>10. first person singular &quot;I&quot;, modal verb, base form verb</td>
<td>personal pronoun &quot;I&quot;, base form verb, singular noun</td>
<td>adjective, adjective singular noun</td>
</tr>
<tr>
<td>11. possessive personal pronoun, singular noun, period</td>
<td>adjective, singular noun, &quot;of&quot;</td>
<td>adjective, singular noun, preposition</td>
</tr>
<tr>
<td>12. adjective, singular noun, period</td>
<td>singular noun, &quot;of&quot;, verb gerund</td>
<td>singular noun, coord conj, singular noun</td>
</tr>
<tr>
<td>13. adjective, singular noun, preposition</td>
<td>period, personal pronoun &quot;I&quot;, base form of &quot;have&quot;</td>
<td>singular noun, singular noun, singular noun, singular noun</td>
</tr>
<tr>
<td>14. singular noun, definite article, preposition</td>
<td>&quot;to&quot;, base form verb, singular noun</td>
<td>singular noun, singular noun, comma</td>
</tr>
<tr>
<td>15. “to”, singular noun, definite article</td>
<td>prep, singular noun, period</td>
<td>singular noun, singular noun, past tense verb</td>
</tr>
<tr>
<td>16. singular noun, coord conj, singular noun</td>
<td>adjective, preposition, singular noun</td>
<td>adjective, plural noun, period</td>
</tr>
<tr>
<td>17. definite article, adjective, singular noun</td>
<td>verb gerund, proper noun, preposition</td>
<td>definite article, adjective, plural noun</td>
</tr>
<tr>
<td>18. present tense verb</td>
<td>def article, singular noun, period</td>
<td>preposition, indefinite article, singular noun</td>
</tr>
<tr>
<td>19. adjective, singular noun, plural noun</td>
<td>&quot;to&quot;, base form verb, adjective</td>
<td>comma, singular noun, singular noun</td>
</tr>
<tr>
<td>20. base form verb, definite article, singular noun</td>
<td>Base form verb, adjective, singular noun</td>
<td>plural noun, preposition, definite article</td>
</tr>
</tbody>
</table>

Four trigrams appeared in both CMC corpora but not in the newspaper corpora. Those were (1) prep-def article-singular noun, (2) personal pronoun "I"-modal verb-basic form verb, (3) adjective-singular noun-period, and (4) present tense verb-definite article-singular noun. The number of appearances of certain parts of speech in each set might also
help to describe each corpus. Articles appear eight times in the collection of native speaker trigrams, and only twice in the non-native speaker data. In non-native speaker data the indefinite article does not appear in any of the most common trigrams.

Trigrams of sequenced emoticons, punctuation with possible parentheses, appeared in the non-native top twenty exclusively. Singular and plural nouns appeared in more than half of the trigrams for each group. Pronouns were not included in these calculations of singular and plural nouns. Singular and plural nouns appeared in all but four of the top twenty native speaker trigrams, totaling 18 appearances altogether. Two of the trigrams had more than one noun. For non-native speakers, nouns appeared less often; they appeared in twelve of the top twenty trigrams and only 14 appearances altogether.

Verbs appeared in 9 of the top twenty native speaker trigrams for a total of 12 appearances altogether. Verbs appeared in 10 of the top twenty trigrams for non-native speakers for a total of 11 appearances altogether. This does not mean native speakers used more verbs than non-native speakers. Similar to previous studies with Asian learners of English, the learner corpus actually contained more verbs, but overwhelmingly in unconjugated forms. Overall verbs made up 12.7% of the non-native sample and 10.6% of the native speaker sample. It is also worth noting that while native speakers used more progressives, gerunds and modal verbs, non-native speakers used almost twice as many base form verbs as native speakers. The newspaper corpus included fewer verbs than either of the discussion board corpora, consisting of only 6.8% verbs and using almost half the percentage of modal verbs as the native speaker sample and an even comparatively lower percentage than that of the non-native speakers. This seems entirely consistent with Heylighen and Dewaele's (2002) findings; we would expect news to be highly context-independent, as compared to bulletin board exchanges, which are more conversational and hence more context-dependent. Not displayed in any of the top twenty trigrams were tags unknown to the tagger which totaled 12% for the native speakers, 16% for the newspaper sample and 19% for the non-native speaker sample.

6. Discussion

The POS tagger revealed clear differences between native and non-native speaker asynchronous CMC. The non-native speaker trigrams mirrored the native speaker trigrams in content but not in frequency. The non-native speaker CMC was comparatively more formulaic—repeating
constructions much more often. Non-native speakers also, consequentially, had fewer unique trigrams. The newspaper sample’s middle point between the two in both the repeated and rarely-appearing POS trigram measures suggests that standard written English is more formulaic than native speaker CMC, but not so much as non-native speaker CMC.

The sample choice, however, may have impacted outcomes. Native speakers were engaged in discussion where they needed to express new ideas and negotiate meaning, while the non-native speakers were trying out their English ability and fulfilling a course task. The range of native speaker POS trigrams may indicate less reporting and more negotiation between the members of the group, and the smaller range found in the non-native speakers’ CMC may have been a result of the tasks’ simplicity.

While these findings may hint at a uniform nature of non-native speaker CMC within this group of Thai learners, that level of speculation was not the point of the study. Rather, uncovering what structural differences a POS tagger might reveal provides insight into the potential gains offered by POS analysis of CMC texts. The differences between these three specific cases, though disparate in their situated nature, does imply that POS analysis can provide a richer description of the structure of native and non-native speaker asynchronous CMC that we would likely be able to guess at by simply reading the texts. While the results of this study are clearly not generalizable to all non-native speakers’ CMC because of the homogeneity of the sample of non-native speakers (and the diversity of the native speakers’ sample) we can imagine that focusing pedagogical activities on the development of unique structures could potentially increase learners’ ability to replicate native-like discourse online. If we decide that native-like discourse is the targeted performance objective, only the development of focused learning materials which address the unique needs of learners and illuminate their learning path can be expected to result in satisfactory performances of learned skills. The non-native writers in this sample, i.e., the Thai students, relied heavily on set phrases or re-used structures transferred from their native language. The POS trigram repetitions of the newspaper corpus, falling between the two bulletin board corpora, hints that this non-native speaker CMC sample is grammatically farther from native speaker CMC than it is from standard written English. Of course, the non-native speakers have a limited lexical and syntactic repertoire which could have impacted the tagger’s ability to tag correctly—the percentage of unknown words was higher in the non-native speaker sample. While it could be the case that unique trigrams in the non-native sample went unnoticed by the tagger due to these unknown tags, it is more likely the case that there actually was more diversity of uniqueness in the native speaker CMC.
However, the grammatical structure of the most frequent CMC trigrams did mirror each other more than the newspaper corpus. The CMC corpora shared four often-used POS trigrams, but none with the newspaper. This may support an argument that the sequences were similar, even if the frequencies were not. The non-native speakers, though producing repetitious structures, may have been aiming at similar expressions unique to CMC, or unique to the task of discussion as opposed to reporting. Because we see the newspaper sample falling between the two bulletin board corpora, we can make the assumption here that situational factors influenced the language more than the mode (Herring 2007).

7. Conclusions and implications for pedagogy

Because the non-native sample relied more heavily on repeated formulaic sequences than either the native or newspaper sample, pedagogical methods of teaching language formulas rather than supporting the creation of unique structures may come into question. In order to bring learners to proficiency in CMC, teachers need to bear in mind that proficient users are creating unique sequences of parts of speech in real time, as they write. This might imply that the pedagogical goal should involve supporting the struggle to negotiate meaning more than to facilitate the production of grammatically common structures. This is not to imply that modeling language items is fruitless, but materials developers would be wise to keep in mind that native speakers do not rely on repeated grammatical structures if findings such as these appear in studies of other groups of non-native CMC as well.

The fact that neither CMC corpus shared especially frequent POS trigrams with the newspaper corpus may suggest that newspaper articles are poor substitutes for reading personal messages if CMC proficiency is a goal. The difference in popular trigrams suggests that a POS tagger could help to describe and distinguish between these types of texts. Materials developers should keep the characteristics of CMC talk in mind when developing curricula. With rich descriptions applied to taggers and natural language processors, there is little reason why a submission text box could not be created which might provide examples of expressions to inform a second language user before posting, in the same sense that grammar checkers provide such options in word processing software now. For example, the overuse of base form verbs presents a teaching opportunity which could be automated using POS tags into an online scaffolding tool. Future research in the proper implementation of these designs, as well as the interplay between mode and situation, would arm the language teacher.
with powerful teaching tools. Similarly, the fact that native speakers used “a” where non-native speakers only used “the” in their most commonly used trigrams suggest that scaffolding software could provide examples with the indefinite article “a” to make learners posts more palatable and better conform to the trends of online discourse. This study provides an example of how the rich descriptions created by tools such as POS taggers expose teaching opportunities we may not have known we had.

REFERENCES


