A corpus-based study on the use of pragmatic markers as speech-like features in Turkish EFL learners’ argumentative essays

M. Pınar Babanoğlu a*

a Süuç İmam University, Aşçar Campus, Kahramanmaraş, 46100, Turkey

Abstract

Authentic learner data is important when investigating the use of the target language by learners during second language acquisition process. Whether it is written or spoken, corpus-based learner data provides an explanation for how learners actually use a foreign language and serves as a relatively narrow empirical basis that second language acquisition research tends to be based on. Analysing learner data has two main goals: to help researchers to better understand the second language acquisition process and to highlight the factors that influence this process (Granger, 2008). Gilquin & Paquot (2007) state that many learners use features more typical of speech than of writing, giving their essays an overly oral tone that may be problematic for learners, as academic writing requires an awareness of the appropriate stylistic use of tone. According to Aijmer (2004), learners may overuse or underuse certain structures in their writing in comparison with native speakers. Pragmatic markers are linguistic features that are more peculiar to conversation than writing, for instance well, you know, I think, etc. and learners may mistakenly include them in their academic writing. This study examines the pragmatic markers in the written text of Turkish English as foreign language (EFL) learners. The study’s aim is to analyse the learners’ use of speech-like features when writing an academic essay. In addition, it is important to investigate whether or not learners overuse or underuse such features when compared to native speakers. Data were gathered from three corpora: the Turkish Corpus of Learner English, the Japanese Corpus of Learner English, and the Louvain Corpus of Native English Essays. Frequency counts and log-likelihood calculations were utilized as quantitative methods to measure the overuse/underuse and to determine whether the possible differences are statistically significant. The results indicated that Turkish EFL learners tend to use oral features in their argumentative essays, which may negatively influence their writing in terms of a stylistically appropriate tone.

© 2014 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).
Selection and peer-review under responsibility of the Organizing Committee of LINELT 2013.

* Corresponding author: M. Pınar Babanoğlu.
E-mail address: pinar.baban@gmail.com
1. Introduction

A corpus is defined as either a collection of texts or a computer database that consists of texts, and is used for the purposes of general linguistics (Meyer, 2002). A learner corpus contains the written or spoken output of foreign language learners. Learner corpora provide an expansive database for analyzing certain linguistic topics in second language acquisition (SLA) and in foreign language learning/teaching research to support L2 investigations. For example, learner corpus data analysis may help to identify some attributes of L2, including:

- linguisitic features in the target language that learners tend to overuse or underuse when compared to native speakers;
- the extent to which learners are influenced by their native language when using the target language;
- learners’ avoidance strategies when they fail to exploit the target language’s expressive possibilities (Leech, 1998).

The study of learner corpora has led to a research program that allows SLA to investigate the above-mentioned issues. The main goal of SLA remains to determine the principles that govern the process of learning a foreign/second language. Granger states that learner corpora contain empirical and natural language data that may be a valuable addition to current SLA data sources (1998). The methodology of learner corpora depends on contrastive interlanguage analysis (CIA). This involves the comparison of learner data with native speaker (NS) data (L2 vs. L1), or the comparison of different types of non-native speaker (NNS) or learner data (L2 vs. L2) (Granger 1996). NS/NNS comparisons are intended to shed light on the non-native features of learner writing and speech through detailed comparisons of the linguistic features in native and non-native corpora. Furthermore, NS/NNS comparisons can highlight a range of non-native features in learner writing and speech, that is, not only errors, but also instances of the underuse and overuse of words, phrases and structures (Granger, 2002). As a sample, the most well-known learner corpus is the International Corpus of Learner English (ICLE) that was composed by Granger et al. (2009). ICLE is a large database of written output produced by learners from 16 different mother tongue backgrounds (Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Turkish, and Tswana) (Granger, et al., 2009). ICLE contains 3.7 million words, all of which have been gathered from the argumentative essays of learners of English from Europe, China, Japan and South Africa. Because of its extensive and rich content, ICLE exhibits a wide range of English as a foreign language (EFL) features.

There is a growing interest in the question of how texts are organized differently depending on whether they are spoken or written. Biber (1988) argues that a variety of techniques can be applied to text corpora in order to identify the underlying dimensions of variation between speech and writing (in Aijmer & Stenström, 2004). Šimčikaitė (2012) claims that the main reason for confusion between spoken and written English is that much more attention has been paid to the grammar of written English, while the grammar of spoken English was not clearly described until recently. In comparison to English, the Turkish language is written as it is spoken though there are some differences between these two genres due to phonological and grammatical structures. A recent study found that Turkish students often transferred, either completely or by paraphrasing, some spoken components when writing, a tendency that is seen as harmful to the language (Kansizoglu, 2012).

In terms of L2 writing, one aspect that may be problematic for learners is the use of an appropriate stylistic tone when writing an academic essay. This is because learners tend to use features that are more typical in speech than in writing (Gilquin & Paquot, 2007). A useful method of analysis of learner writing and speech involves the utilization of corpus data as well as drawing comparisons between native and non-native speakers. This study uses corpus-based research to investigate the use of oral features in EFL learners’ argumentative essays, specifically pragmatic markers such as well, you know, and I mean.

1.1 Pragmatic markers

An important point to highlight is that conversation is different from both writing and formal speech. Conversation is unplanned and is produced under cognitive constraints that are expressed by filled and unfilled pauses, repetition and incomplete grammatical structures (Aijmer, 2004). Lexical items such as well, you know, you see, actually, and sort of, are more peculiar to spoken language and have been called “pragmatic markers” by
Aijmer (2004). Shiffrin (1987) considers pragmatic markers to be “discourse markers” and characterizes them as deictic as well as suggesting that they have identical functions due to different discourse planes. According to O’Keefe et al. (2007), discourse markers are frequent in spoken vocabulary and function by organizing and monitoring the progress of speech. The most common pragmatic markers in the mostly used top 2000 words are *I mean, right, well, so, good,* and *anyway,* all of which function when opening and closing conversation, as well as when speakers return to a topic after an interruption. Aijmer & Simon-Vandenbergen state that a “Pragmatic marker is preferred to discourse markers when the markers have a pragmatic rather than a discourse-marking function” (2006, p. 9). They also argue that the term pragmatic marker is used more generally to refer to a wide variety of both interpersonal and textual functions.

Traditionally, writing was considered a carefully planned activity and thus, it was not likely to entail the use of pragmatic markers that are seen as features of conversation. The distinction, however, between speech and writing has blurred following recent advancements in communication technologies; written communication such as email, instant messaging and texting is much closer to speech than earlier forms of writing and has brought with it “a wealth of written discourse marker use” because users of these forms of communication tend to treat the interactions rather like a conversation (Fox Tale, 2010).

In terms of linguistics, pragmatic markers may contain elements such as main clauses (*I mean, sentence adverbials (actually, frankly, so)*) and conjunctions (*and*). Whether at a textual or an interpersonal level, words and phrases that function as pragmatic markers are often ambiguous because some of them share both marker function and adverbial function, such as *well* and *now.* In relation to this ambiguity, the categorization of pragmatic markers is debated, for example, Hansen describes how the establishing of taxonomy for markers is problematic: “…. compared many other areas within linguistic, the study of markers is relatively recent phenomenon, and attempting an exhaustive taxonomy of content categories in this domain….simply seems premature as long as there is little consensus both about the function of individual morphemes, and about exactly which items should be included in the class of markers.” (Hansen, 1998, in Aijmer, 2002, p.38)

More specifically, *well* is a versatile discourse marker but it functions generally as a “deliberation signal,” reflecting the speaker’s need to give a brief thought or consideration about the point at issue (Biber, et al., 1999). *Now* is an utterance launcher that function as “a bit of conversational space,” marking either a return to the related subject or a new departure, as well as providing additional background information and continuing the current topic of the conversation. Additionally, other markers that have similar functions, like *I mean,* are used to clarify something, while *you know* is used to assume shared knowledge, whereas *I think* and *I guess* express uncertainty, also sort of signals fuzziness, *though* expresses concession, and *and* is used to add information. However, when *though* is placed in a sentence-final position and when *and* is placed in a sentence-initial position, they are considered to be more characteristic of speech rather than writing (Gilquin & Paquot, 2007).

### 1.1.1 Pragmatic markers in L2

The study of pragmatics has recently entered fields such as SLA and has been termed “interlanguage pragmatics” by Aijmer & Simon-Vandenbergen (2006). Pragmatic markers are useful language devices for learners’ communicative needs. When conversing with native speakers, the level of communicative stress that learners experience is often high, though this may be controlled by the use of markers (Aijmer, 2004). Markers help learners to negotiate their way through unplanned conversation. That said, L2 learner attitudes towards the use of pragmatic markers in writing is accepted as problematic because of the prevalence of these markers in academic writing gives their essays an oral tone and cause stylistical inappropriateness (Gilquin & Paquot, 2007).

By means of learner corpora, many researchers have investigated pragmatic markers in L2 English by focusing on the different aspects of the markers, including, Granger and Ryson (1998), Granger (1998), Altenberg and Tapper (1998), Aijmer (2002, 2004), Müller (2005), Narita and Sugiura (2006), Gilquin and Paquot, (2007), Jabeen et al. (2011), and Simcikaite (2012). Many of these studies show that oral features in learner essays tend to be overused when compared to the writing of native speakers. A limited number of L1 speakers also adopt an oral tone in their writing, probably because of a confusion of register (Crawford, 2005, in Gilquin & Paquot, 2007). In the same way, the cause-effect adverbial *so* has been taken into consideration in terms of its specific patterning with particular items such as *so on* and *or so* that are common in spoken language.
1.2 Research Questions

In this study, pragmatic markers, or “speech-like features,” are investigated by means of a learner corpus methodology. The research questions for the study are as follows:

1. Is there a statistically significant difference between native English speakers and EFL learners in the use of oral features in their argumentative essays?
2. Is there a statistically significant difference between two EFL learner groups in the use of oral features in their argumentative essays?

2. Methodology

This study will utilize learner data from the Turkish Corpus of Learner English (TICLE) and the Japanese Corpus of Learner English (JPICLE), both of which are sub-corpora of the International Corpus of Learner English (ICLE) database. The ICLE corpus consists of argumentative essays from EFL learners from a variety of backgrounds. The Louvain Corpus of Native English Essays (LOCNESS) is a reference corpus of native English and consists of American university students’ argumentative essays. Table 1 shows the number of words and essays for each of the three corpora:

<table>
<thead>
<tr>
<th></th>
<th>TICLE</th>
<th>JPICLE</th>
<th>LOCNESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of words</td>
<td>169345</td>
<td>168160</td>
<td>168325</td>
</tr>
<tr>
<td>Number of essays</td>
<td>280</td>
<td>315</td>
<td>207</td>
</tr>
</tbody>
</table>

The analysis to examine the pragmatic markers within three corpora as TICLE and JPICLE as learner data sources and LOCNESS as native data was carried out by Wordsmith tools (Scott, 2008). The concordance facility of Wordsmith tools helps to identify the pragmatic structures within a corpus. Following this, frequency calculations of oral features will be gathered from each of the three databases and will be compared using CIA. Specifically, the frequency of markers will be compared between the TICLE and the LOCNESS corpora and between the JPICLE and the LOCNESS corpora in terms of L2 vs. L1 (between a learner group and a native group), as well as between the TICLE and the JPICLE databases in terms of L2 vs. L2 (between two learner groups). Frequency differences are measured by a log-likelihood (LL) statistical measurement. Log-likelihood (LL) ratio is a sophisticated type of statistical measurement that is utilized especially in corpora comparisons and calculates frequencies by considering the word sizes of two corpora as well as determining the relative overuse and underuse in a certain corpus.

3. Results

The speech-like pragmatic markers used by Turkish and Japanese learners and native English speakers in their argumentative essays are listed in Table 2, as shown below:

<table>
<thead>
<tr>
<th>Oral features</th>
<th>TICLE f</th>
<th>JPICLE f</th>
<th>LOCNESS f</th>
</tr>
</thead>
<tbody>
<tr>
<td>I think</td>
<td>174</td>
<td>732</td>
<td>41</td>
</tr>
<tr>
<td>and (sentence initial)</td>
<td>193</td>
<td>408</td>
<td>37</td>
</tr>
<tr>
<td>(and) so on</td>
<td>21</td>
<td>111</td>
<td>4</td>
</tr>
<tr>
<td>kind of</td>
<td>127</td>
<td>72</td>
<td>14</td>
</tr>
<tr>
<td>Maybe</td>
<td>78</td>
<td>61</td>
<td>31</td>
</tr>
<tr>
<td>Actually</td>
<td>26</td>
<td>73</td>
<td>38</td>
</tr>
<tr>
<td>like that</td>
<td>22</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>Well</td>
<td>6</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>Thanks to</td>
<td>25</td>
<td>15</td>
<td>3</td>
</tr>
</tbody>
</table>
Overall, the results showed frequency differences both between NNS and NS and between NNS and NNS. Japanese learners seemed to use the highest amount (1615) of pragmatic markers whereas Turkish learners used about half of this amount (814), though still considerably more than native English speakers (228). Table 3 shows this in relation to the three corpora:

Table 3. Overall frequency of markers among three corpora

<table>
<thead>
<tr>
<th>TICLE</th>
<th>JPICLE</th>
<th>LOCNESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Markers</td>
<td>814</td>
<td>1615</td>
</tr>
</tbody>
</table>

Nevertheless, a statistical measurement is needed to demonstrate the significant difference between the groups. Table 4 shows the LL measurements for the overall frequencies among the three corpora:

Table 4. Log-likelihood ratio (LL) of overall frequency of pragmatic markers among learner and native corpora

<table>
<thead>
<tr>
<th>TICLE vs. LOCNESS</th>
<th>JPICLE vs. LOCNESS</th>
<th>TICLE vs. JPICLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LL</td>
<td>LL</td>
<td>LL</td>
</tr>
<tr>
<td>Markers</td>
<td>346.06+</td>
<td>1176.80+</td>
</tr>
</tbody>
</table>

$p < 0.05$ (critical value: 3.84); + indicates overuse in the first corpus relative to the second corpus; - indicates underuse in the first corpus relative to the second corpus.

The LL results confirmed the possible overuse of pragmatic markers in Japanese data both against native and Turkish speaker data. If the critical value of LL is considered as 3.84, 1176.80+ by JPICLE over LOCNESS reveals a fairly significant difference between Japanese and native English speakers in the use of pragmatic markers. Similarly, when Turkish learner data is compared with native speakers, the 346.06+ overuse is a considerable difference between the Turkish learner group and the native English speakers. That said, since this study is concerned with Turkish learners, the comparison between the two learner groups, TICLE and JPICLE, revealed underuse in Turkish learners when compared with Japanese learners. This fact may be a sign of a language-specific reason since there is no shared behaviour between learner groups in the use of markers. However, it is clear that, both learner groups showed overuse in the use of pragmatic markers against native English speakers.

The high frequency in the JPICLE corpus seems to be mostly because of the use of certain pragmatic particles, such as *I think* (732), the sentence-initial *and* (408), and *so on* (111). Interestingly, *I think* was the most frequent marker in both the Turkish learner data (174) and the native speaker data (41) though both of these figures fall far
below Japanese learners. Table 5 presents the frequency of I think:

Table 5. Frequency of I think in Turkish, Japanese and Native English Speaker Writing

<table>
<thead>
<tr>
<th></th>
<th>TICLE</th>
<th>JPICLE</th>
<th>LOCNESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>I Think</td>
<td>174</td>
<td>732</td>
<td>41</td>
</tr>
</tbody>
</table>

I think is accepted as rare in academic writing (54.38 occurrences per million words) (Gilquin & Paquot). In this study, it revealed considerable frequencies in learner writing (10.2 per 10,000 words in TICLE and 43.5 per 10,000 words in JPICLE). Samples from TICLE and JPICLE can be seen below:

1. .....it is a plague for the teachers. I think that there are not any students who have not cheated..(TICLE/TRCU.1004.txt)
2. .....so we don’t actually feel the need to put them in practice. So, I think we should begin ...(JPICLE/JPKO.2023.txt)

The markers as and (sentence initial), kind of, maybe, I mean are other frequently used markers in TICLE and JPICLE illustrated in Table 4. :

Table 6. Type/token ratio of and (sentence initial), kind of, maybe, I mean in three corpora

<table>
<thead>
<tr>
<th>Oral Features</th>
<th>f</th>
<th>TICLE T/t</th>
<th>JPICLE T/t</th>
<th>LOCNESS T/t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maybe</td>
<td>78</td>
<td>9.5</td>
<td>61</td>
<td>31</td>
</tr>
<tr>
<td>I mean</td>
<td>35</td>
<td>4.2</td>
<td>9</td>
<td>0.5</td>
</tr>
<tr>
<td>kind of</td>
<td>127</td>
<td>15.6</td>
<td>72</td>
<td>14</td>
</tr>
<tr>
<td>And (sentence initial)</td>
<td>193</td>
<td>23.7</td>
<td>408</td>
<td>37</td>
</tr>
</tbody>
</table>

Type/token ratio results show that, markers I mean and kind of seem to be used by Turkish learners more than both Japanese and native English speakers whereas Maybe has the highest rank in native English data although its the lowest frequency. On the other hand, sentence-initial and is one of the mostly used markers by Japanese learners and also by Turkish learners. In fact, and is commonly used as a conjunction in academic writing but is assumed to be rare in a sentence-initial position that is characteristic of speech rather than of academic writing. The sentence-initial and occurred 11.3 times per 10,000 words in TICLE and 24.2 per 10,000 words in JPICLE and are shown in the following extracts from learner corpora:

3. people force to change their lifestyle and places where they live. And also, when the people do not care on holdovers... (TICLE/TRCU.1133.txt)
4. family environment will be complicated. And more serious problem is that... (JPICLE/JPTM.1016.txt)

Results indicated considerable overuse by learner groups against native speakers of English. Although two learner groups differs in the choice of markers, a common attitude can be noticed in the tendency of using them in general.

4. Discussion and conclusion

Considering the frequency calculations and log-likelihood results, Turkish and Japanese EFL learners when compared to native English speakers overuse certain pragmatic markers. This result is not surprising since previous
studies indicated similar results (Gilquin & Paquot, 2007; Simcikaite, 2012). As emphasized by colleagues who found similar results, the overuse of these structures by EFL learners is due to various reasons, such as register confusion, L1 transfer, L2 instruction, as well as developmental factors (Gilquin & Paquot, 2007). In terms of Turkish learners, it can be said that the overuse of oral features in writing is related to the current standards of L2 instruction that are based on a communicative approach. Course textbooks that take a communicative approach as the major theoretical background for grammar instruction may promote confusion between written and spoken registers among learners by focusing on a communication-oriented task. Correlatively, learners may over generalize the communicative aspects in their writing by using lexical items or expressions in inappropriate contexts to overcome the difficulties of using the target language. Another point for consideration relates to L1 transfer, a phenomenon that was observed in Turkish learners because markers like I think (bence, sanırım), maybe (belki), I mean (yani), kind of (bir nevi, bir tür) are very common in both spoken and written Turkish. According to Dursunoğlu (2006), in Turkish, the difference between written and spoken language is negligible, which is unlike many other languages. This may be one of the reasons for Turkish learners’ overuse of oral features in their writing, however, this suggestion would require a detailed analysis in order to state an L1 transfer, at least to investigate the use of markers in Turkish language in terms of comparison.

Among the learner groups, Japanese learners used pragmatic markers more than Turkish learners. The overuse of markers in Japanese learners was remarkable and may be due to an L1 transfer, which means that all learners did not share a language-specific condition. When considered against native speakers from previous studies, both Turkish and Japanese learners tend to use more oral features in their essays than native speakers.

The outcome of this study can be regarded as an indicator for an EFL methodology that might help learners to use a stylistically appropriate tone in their L2 writing. Consciousness rising and/or awareness in the genre differences as writing/speech differences might be helpful for learners to product stylistically appropriate tone in their academic essays. Highlighting cross-linguistic differences between the native language and the target language as well as providing relevant L2 writing instructions may also support learners. If learners are encouraged to focus on a more formal style of academic writing they may become more aware of writing styles. Another suggestion is to include corpus-based techniques for learning contexts. For instance, the study of corpora like the British National Corpus, which comprises a 100 million word collection of samples of British English including academic books, scientific periodicals, journals, and school and university essays written by students with data-driven learning methods may encourage researchers to explore authentic data and to determine the formal styles found in written forms of British English.

For further research, considering L1 transfer, especially since Japanese learners overuse markers far more than Turkish learners may focus on the significant overuse of pragmatic markers in Turkish and Japanese learners’ essays. Further research might also include data from other learner groups from different L1 backgrounds in order to clarify the differences in pragmatic marker usage by comparing the performances of various learner groups.

References


