saga
Vol. 1 No. 1, 2020
P-ISSN: 2715-7512; E-ISSN: 2716-0246
doi: http://doi.org/10.21460/saga.2020.11.28

INTEGRATING LEARNER CORPUS ANALYSIS INTO THE TEACHING OF ENGLISH ACADEMIC WRITING

Lavinia Disa Winona Araminta
Universitas Indonesia
laviniadisa@ui.ac.id

Received: 28 November 2019 Accepted: 12 December 2019

Abstract

Practical implementation of learner corpus research to inform language pedagogy has been common, given the availability of resources, such as a large amount of data about the products of language learning and factual language uses, and the necessary technology, such as concordance programs. This article lays out the typical analyses of learner corpora and the implications of and issues surrounding such studies on second/foreign language teaching based on the existing literature. More specifically, the article captures the need for a more extensive corpus of Indonesian learners' English other than what is already available to represent more insights about English language teaching in Indonesia. Furthermore, it proposes the development of an in-house learner corpus for direct and indirect uses at Universitas Indonesia. An actual trial on building a sample learner corpus and running a lexical analysis demonstrates the plausibility of integrating learner corpus into the teaching of academic writing on higher-education levels.

Keywords: Academic writing; English language teaching; learner corpus

INTRODUCTION

A learner corpus is generally defined as a collection of texts produced by learners of a particular language (Hunston, 2002). With technological advancement, the compiling and storing of a learner corpus can be computerized and done in large quantities, and the analysis of it can be made automated (Granger, 2002). A later definition regards a learner corpus as an electronic collection of foreign or second language learner texts which are assembled based on explicit design criteria (Granger, 2009).

Around the world, there have existed more than 150 learner corpora with different target (L2) and first languages (L1), mediums, text types, task types, proficiency levels, and sizes in words (Université catholique de Louvain, 2017). This variety is attributed to the explicit design

criteria of each learner corpus, making it distinct from other learner corpora and specifying the characteristics of the corpus. For example, in terms of first languages (learner variable) and genres (task variable), the International Corpus of Learner English (ICLE) differs from the British Academic Written English (BAWE) corpus. The former collected texts from English as a Foreign Language (EFL) learners with 11 different European mother tongue backgrounds and focuses on essay writing (Granger, 2003). Meanwhile, the latter gathered texts mostly written by native speakers of English and covers 13 broad genre families, including essays, case studies, and methodology recounts (Hyland, 2008). Nevertheless, in terms of medium, both corpora consist of written, instead of spoken, texts.

In addition to explicit design criteria, the other key features of a learner corpus include: 1) being gathered from genuine communicative events or authentic classroom activities; 2) being situated in non-native, including FL (foreign language) and SL (second language), varieties of the target language; 3) consisting of continuous stretches of discourse instead of isolated words or sentences; 4) being collected for a particular SLA (Second Language Acquisition)/ FLT (Foreign Language Teaching) purpose, and; 5) if the corpus is annotated, following a certain standard and being documented for learner and task variables (Granger, 2002).

DISCUSSIONS

Learner corpus analyses

One methodological approach to linguistic analysis of learner corpora is comparative, which is to identify the extent to which learners or non-native speakers (NNS) differ from each other and from native speakers (NS) with regard to the language they produce (Hunston, 2002). To achieve this purpose, a comparison between two comparable corpora is essential, such as between ICLE and the Louvain Corpus of Native English Essays (LOCNESS) whose texts are of the same genre—essay—but were produced by writers of different native languages. Such a comparison is termed Contrastive Interlanguage Analysis (CIA). It involves two types of comparisons.

The first type is NS/NNS comparisons, which can pinpoint non-native features of learner writing and speech by comparing non-native to native learner corpora (Granger, 2002). What are frequently found in this type of comparison are instances of overrepresentation and underrepresentation of words, phrases, and structures. For example, a comparison between the Swedish sub-corpus of ICLE and LOCNESS revealed differences between NNS and NS in terms of how they organized information in argumentative writing (Herriman & Aronsson, 2009). Using a concordance program, the study found that the NNS overused certain types of themes and thematic variation, such as subjective interpersonal metaphors (*I think*) and conjunctive textual themes (*however*). These features, according to the researchers, made their style of persuasion more emphatic and their style of writing fit spoken language, rather than written one.

The second type is NNS/NNS comparisons, which can further investigate interlanguage in SLA by comparing two or more non-native learner corpora from different L1s (Granger, 2002). One assumption generated from this type of comparison is that linguistic features shared by several learner populations are more likely to be developmental and those found only in the data from one national group may be subject to the learners' L1 (Granger, 2002, p. 13). To illustrate this, a study comparing German and Italian students' writings in ICLE found that Italian learners overused more text connectors than German advanced learners did (Waibel, 2005). The results were then compared to Granger and Tyson's (1996) study of Italian and French learners in

Waibel's (2005). It was concluded that German learners were generally more proficient than Italian and, mostly, French learners in using text connectors. The researcher suggested that L1 transfer was the possible cause for a few instances of under- and overused structures, but language universals were unlikely the case. Instead, learners' unawareness of NS usage and the different teaching methods in the respective countries might have contributed to the findings.

Another approach to analyzing learner corpus is computer-aided error analysis (CEA), employing computer tools to tag, retrieve, and analyze errors (Granger, 2002). With a raw-text corpus, error-prone linguistic items can be pre-selected and scanned in the corpus to find instances of misuse. For example, a study using the German sub-corpus of ICLE found that one fourth to one fifth of the use of support verb constructions, such as *make changes* and *have a look at*, by advanced German-speaking learners was wrong. The identified mistakes included wrong verb, wrong verb and noun, and wrong determiner (Nesselhauf, 2004).

The second option in CEA is tagging a learner corpus for all errors or errors in particular categories based on a standardized system of error tags (Granger, 2002). For example, to investigate the issue of second language accuracy developmental trajectories, the German, French, and Spanish components of ICLE were manually error-tagged according to the Louvain error-tagging taxonomy which covers seven main error domains, such as grammatical, lexicogrammatical, and style errors. The tagging resulted in 45 error types, each of which was counted for its occurrence at each level of proficiency to find points of progression, stabilization, and regression (Thewissen, 2013). Although this process is more labor-intensive, the search for errors can be expanded, instead of being limited to a certain pre-selected linguistic item (Granger, 2002).

Besides error tagging, another way to annotate a learner corpus is part-of-speech (POS) tagging, which can inform the word-class membership of each word in a corpus (Granger, 2002) and, thus, makes it easier to do an extraction of words belonging to particular parts of speech. One example is a study examining L1 influence on the acquisition order of English grammatical morphemes by L1 Japanese, Korean, Spanish, Russian, Turkish, German, and French learners of English from five proficiency levels (Murakami & Alexopoulou, 2016). It used the Cambridge Learner Corpus (CLC) which contained both parts of speech and grammatical relations, and focused on six most frequently studied morphemes including the past tense *-ed* and plural *-s*. Since the corpus was also error-tagged, accuracy scores for the use of the target morphemes could be obtained, revealing differences in the accuracy order across proficiency levels as well as across and within L1s.

While most of the studies reported here are cross-sectional and portray the characteristics of certain groups of learners at one single time, a couple of them are quasi-longitudinal (Murakami & Alexopoulou, 2016; Thewissen, 2013). Longitudinal studies are possible to carry out, but it needs a longitudinal learner corpus as well. This kind of corpus can be developed for research purposes, for instance, by collecting essays written by two L2 German learners over four consecutive semesters to investigate the development of their writing complexity (Vyatkina, 2012). The reported studies here also show that multiple approaches can be employed in one study, such as the combination of error tagging and NNS/NNS comparisons (Thewissen, 2013). Moreover, comparisons of learner corpora can be conducted not only based on L1 and level of proficiency of the learners. For instance, in the context of English for Specific Purposes (ESP) situated in Hong Kong, an apprentice and a professional corpus of technical recommendation-type reports were compared to uncover the lexis for the Problem-Solution pattern in each corpus (Flowerdew, 2004).

Pedagogical implications and issues to consider

Since a learner corpus is collected for SLA or FLT purposes, the results of learner corpus research are likely to have pedagogical implications, albeit to various extents. In terms of who can benefit from the results, Granger (2009) distinguished between delayed and immediate pedagogical use of learner corpora. The former is usually larger and has wider generalizability to similar-type learners. On the other hand, the latter is smaller, represents a more specific learner population and variety of language and, thus, is more relevant to be applied in the classroom.

Nevertheless, the above distinction should not be seen as a clear-cut division but, instead, two ends of a continuum. Studies on interlanguage and the development of learner language, such as Vyatkina's (2012), are closer to delayed pedagogical use since they typically deal with SLA rather than practical issues in FLT. Cross-sectional studies, such as Nesselhauf's (2004), are in between since if learners' proficiency is increasing along with the improved teaching practices in the specified contexts, the results of these studies may no longer be relevant. At the end of immediate pedagogical use are the cases in which learners are engaged with their own productions (Granger, 2002) or what is termed as 'learning-driven data', which lets learners be the researchers (Seidlhofer, 2002). It is important to note that the learners in Seidlhofer's study were future teachers of English and likely to benefit from using a concordance program. However, this teaching approach may not be practical and less relevant to other groups of learners.

In terms of improving classroom practices, there is a warning against directly translating the results of learner corpus analyses into teaching recommendations (Granger, 2009; Hunston, 2002). To avoid giving misleading advice, teachers need to critically interpret the results of comparison-type studies. For example, an overuse of particular words by NNS does not necessarily mean that learners should use those words less often. Rather, a further investigation needs to be conducted to know the circumstances when NS would typically use alternative words and what the alternative words are (Hunston, 2002). Another example is the suggestion to teach learners the cultural norm differences of argumentative writing in their L1 (Swedish) and the target language (English) (Herriman & Aronsson, 2009). This pedagogical practice can be useful for learners, especially if L1 transfer is found to be the possible cause of their overuse of certain linguistic items.

The results of CEA do not need to be attended to one by one. The analyses should not be aimed at eliminating as many errors as possible but drawing conclusions on which linguistic items or structures appear to be the most difficult to understand and produce by learners. These can lead to suggestions pertaining to the order of what to teach. One way to do this is by looking at the frequency of error types. The most frequently occurring error type can be assumed to be the most difficult item for learners and, thus, should be taught first. For instance, in teaching support verb constructions to advanced German-speaking learners of English, choosing the right verbs was suggested to be the first focus of teaching, followed by choosing the right noun complementation as well as the right noun and then contrasting verb constructions with similar verbs (Nesselhauf, 2004).

Another way of doing a difficulty-ordering is having learners' levels of proficiency identified. Using the WriCLE corpus and the UPV Learner Corpus, O'Donnell (2015) identified three general patterns of changing usage of linguistic features in Spanish university students' English. Increasing usage that was in line with increasing proficiency could include a feature which was not part of L1 but needed to be acquired. Decreasing usage that was opposed to increasing

proficiency may include a feature transferred from L1, which in a later stage was not used anymore as learners became more proficient. Initially rising usage which then decreased might refer to a feature that learners had difficulties with at first but was later overcome as they gained in proficiency. The results of such a study can suggest the order of what to teach not only, for example, in one semester but also over five semesters.

Possible implementations at Universitas Indonesia

At a glance, the studies of learner corpora that I have come across or at least that are reported in this paper tend to 'overuse' ICLE. The fact that only few learner corpora are available for public use (Granger, 2002; Waibel, 2005) may explain why ICLE, being a large learner corpus published on CD-ROM, is frequently researched into. Although it is possible to draw more reliable conclusions about learner language (Waibel, 2005), the results of studies using ICLE may not be generalizable to Indonesian learners of English. To date, the International Corpus Network of Asian Learners of English (ICNALE) is the only corpus known for containing a component of Indonesian learners (Ishikawa, 2013). This corpus is publicly available, but, being a large learner corpus, its pedagogical use is somewhere between delayed and immediate. Meanwhile, with the availability of technology and the Internet, Granger (2002) encouraged the collection of smaller in-house corpora, for instance, by collecting soft copies of students' works via email. The pedagogical use of this kind of corpora can apparently be more immediate than that of larger learner corpora, and information resulted from it will be valuable for making specific suggestions on teaching practices, material development and evaluation in a specific institution.

For those reasons, building an in-house learner corpus of English from L1 Indonesian learners seems to be plausible. To try out this idea, I would like to propose developing one from data about learners of English studying at Universitas Indonesia. The corpus will: 1) consist of texts written by first-year students during authentic classroom activities in an English for Academic Purposes (EAP) course; 2) focus on FL varieties of English; 3) contain textual data, and; 4) be collected for FLT purposes. The next things to consider are standardization—if the corpus is to be error- and POS-tagged—and documentation. Regarding its explicit design criteria, most learners are L1 Indonesian learners, except those who speak local languages as their L1s. Their levels of proficiency vary, and this information can be obtained from the results of English Placement Test (EPT) they do during the orientation for new students. The texts they write fall into two types, which are 150-word article summaries and 750-word essays collected from classroom assessments and final term tests. The teacher usually determines the articles for the summary. While for the essay, the students can choose any topics or one out of five topics provided in the final test. In the former case, essays can range from expository to argumentative writing.

To illustrate the implementation of learner corpus analysis at Universitas Indonesia, I compiled 20 essays on various topics submitted for classroom assignments via email by my former students who majored in Computer Sciences. Using regular expressions on MonoConc Pro, I searched for the lemma "make" and found that this lemma was commonly followed by:

1. Object (single noun/noun phrase)

```
20. ... on making skill. Children will learn to [[make]] decision from simple thing like how to  ...
23. ... volunteering can be a meaningful way to [[make]] new friends. Networking is an exciting  ..
 5.  ...  texts and never as commands. Thirdly, [[make]] good use of server-side validation. Cli ...
```

2. Object + complement (adj.)

```
29. ...  warm themself. Other than that alcohol [[makes]] your mind free, like you didn't feel st ...
 9. ...  need to diversify our energy source to [[make]] it more sustainable. There are many rea ...
21. ...  self-confidence, and self-esteem which [[make]] them ready to face anything, including  ...
```

3. Object + bare infinitive

```
15. ... tivities to relieve their stress and to [[make]] them feel relaxed. This due to the beli ...
16. ... lso known as an addictive game that can [[make]] students become an all-nighter. It can  ...
27. ... ey, cogna, wine and many more. And what [[makes]] people always come back to drink that a ...
```

There was also one idiom found:

```
8. ... ill always find a way to get around and [[make]] their way in. Because of that, broad kn
```

Based on the abovementioned lexical analysis, a number of ideas for teaching academic writing in English can be implemented. For immediate pedagogical use, good examples of clause patterns and idioms can be shown to students as models in teaching grammar for writing. Wrong collocations, such as noun complements, can also be pointed out. Students can be invited to discuss why they choose wrong collocations, what the consequences are (e.g. readers will not understand), whether they need improvement, and what kind of improvement they need. This can be a point of departure for teachers to develop or improve materials and/or devise activities tailored to learners' needs.

For less immediate pedagogical use, the teachers can try comparing learners' classroom assignments to final term tests in terms of accuracy/errors, complexity, and fluency to inform learners' progress. The summary of the progress can be reported to learners at the end of the course with regard to areas that can be improved. For delayed use, the corpus can be compared to both ICNALE and BAWE to identify the features of students' writing at Universitas Indonesia which are different from those of other Asian learners of English and native English writers. However, since the sample corpus is not annotated, the search was limited to a specific lemma. Having the corpus error- and POS-tagged will expand the search and findings.

CONCLUSION

The literature review presented in this study leads to several takeaways. First of all, learner corpora have been valuable for language teachers and researchers in, among others, analyzing learners' major weaknesses and areas for improvement, identifying certain errors which learners frequently make, and monitoring their learning progress. Second, learner corpus analyses can be useful for either direct classroom applications or further studies into a group of language learners or those of shared backgrounds. Nevertheless, to avoid hasty conclusions, any results obtained from a learner corpus analysis have to be interpreted critically by considering the characteristics of both learners and data as well as the design of the analysis.

Despite the availability of learner corpora from various countries and in different forms, data about certain groups of learners remain underrepresented. In Indonesian context, there is only one corpus containing data on Indonesian learners of English. To gain more understanding about Indonesian learners, more data are needed. As building a wide-ranging learner corpus may be costly and time-consuming, a more practical option is to develop an in-house learner corpus. A sample corpus was gathered from a selection of essays written by students at Universitas Indonesia. The results of the lexical analysis lead to some pedagogical ideas and show that building an in-house learner corpus in a specific institution is realistic and may even be of

great benefit for language teachers, curriculum designers, and course developers due to its data-driven approach to understanding the nature of learners' language learning.

REFERENCES

Flowerdew, L. (2004). The problem-solution pattern in apprentice vs. professional technical writing: An application of appraisal theory. In G. Aston, S. Bernardini, & D. Stewart (Eds.), *Corpora and language learners*, (pp. 125-135). Philadelphia, USA: John Benjamins Publishing Company.

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching*, (pp. 3-33). Philadelphia, USA: John Benjamins Publishing Company.

Granger, S. (2003). The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly, 37*(3), 538-546.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Ajimer (Ed.), *Corpora and language teaching*, (pp. 13-32). Philadelphia, USA: John Benjamins Publishing Company.

Herriman, J. & Aronsson, M. B. (2009). Themes in Swedish advanced learners' writing in English. In K. Ajimer (Ed.), *Corpora and language teaching*, (pp. 101-120). Philadelphia, USA: John Benjamins Publishing Company.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge, United Kingdom: Cambridge University Press.

Hyland, K. (2008). The British Academic Written English (BAWE) corpus. *Journal of English for Academic Purposes, 7*, 294.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner Corpus Studies in Asia and the World, 1,* 91-118.

Murakami, A., & Alexopoulou, T. (2016). L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition, 38*(3), 365–401. doi: 10.1017/S0272263115000352

Nesselhauf, N. (2004). How learner corpus analysis can contribute to language teaching: A study of support verb constructions. In G. Aston, S. Bernardini, & D. Stewart (Eds.), *Corpora and language learners*, (pp. 109-124). Philadelphia, PA: John Benjamins Publishing Company.

O'Donnell, M. (2015). Using learner corpora to order linguistic structures in terms of apparent difficulty. In E. Castello, K. Ackerley, & F. Coccetta, *Studies in learner corpus linguistics: Research and applications for foreign language teaching and assessment*, (1st ed., pp. 71-85). New York, NY: Peter Lang.

Seidlhofer, B. (2002). Pedagogy and local learner corpora: working with learning-driven data. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching*, (pp. 213-234). Philadelphia, USA: John Benjamins Publishing Company.

Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal, 97*(S1), 77–101. doi: 10.1111/j.1540-4781.2012.01422.x

Université catholique de Louvain. (2017). "Learner corpora around the world". https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal, 96*(4), 576–598. doi: 10.1111/j.1540-4781.2012.01401.x

Waibel, B. (2005). Corpus-based approaches to learner interlanguage: Case studies based on the "International Corpus of Learner English". *AAA: Arbeiten aus Anglistik und Amerikanistik, 30*(1/2), 143-176.