ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 13, No. 2, pp. 417-423, February 2023
DOI: https://doi.org/10.17507/tpls.1302.16

A Review of Research on Learner Corpora —Taking Overseas Core Journals in Linguistics From 2007 to 2021 as an Example

Shune Yang
Shanghai International Studies Universities, Shanghai, China, 200083

***Abstract*—To gain an in-depth understanding of the development status and trend of research on learner corpora, the study, employing the Web of Science (WoB) as a data source, conducts a statistical analysis of research on learner corpora published in internationally renowned academic journals from 2007 to 2021, focusing on the respects of research trends, disciplines, journals, language, and content, and others. Results showed that the international research on learner corpora is characterized by an interdisciplinary tendency, diverse topics, single language, and uneven distribution. Based on these features, relevant suggestions are made for both learners and teachers.**

***Index Terms*—corpus, corpus linguistics, learner corpora, research on learner corpora, foreign language teaching**

I. Research Background

Learner corpus, a branch of corpus linguistics, is an emerging field in second language acquisition research (Granger, 2004). Unlike general corpora, learner corpora refer to the computerized electronic collections of the language output produced by foreign language learners, known as interlanguage (Leech & Garside, 1998). Learner corpora, employing the main principles, tools, and methods of corpus linguistics, provide the basis for analyzing the output of learners' language. This aids the analysis of learners' error characteristics and overall language characteristics (Li, 1999), which opens up research ideas and methods for the field of second language acquisition. Pravec (2002) pointed out in his "Survey of Learner Corpora" that the establishment of learner corpora can collect objective data to describe learners' language. Analyzing these data will enable researchers to focus on the issues of language theory and teaching practice and teachers to focus on the real needs of learners. In addition, this study provides detailed information on numerous existing and easily accessible learner corpora, which is very convenient for researchers to use in language study, teaching, and learning.

Previous research on second language acquisition mainly relies on introspective methods or other methods like questionnaires, which obtain the required data by induction. However, such data provided by participants could be artificial and susceptible to subjective factors, which may limit the generality of the data. Accordingly, the representativeness of the obtained data might affect the conclusion of these studies. Callies (2015) stated: "compared to other types of data traditionally used in second language acquisition research, learner corpora provide the authentic, continuous, and contextualized systematic collection of the language of foreign language learners, which is stored in an electronic format" (p. 35). Obviously, learner corpora provide a large and objective database platform for second language acquisition research, which effectively reinforces the lack of data sources in traditional research on second language acquisition. As a research method, learner corpora, in addition to being used to test research hypotheses, could also be used to generate new hypotheses. Therefore, instead of proposing research topics in a preconceived way, researchers could use software to evaluate and analyze data to determine what patterns and problems may exist in learners' language, which helps to generate new and valuable topics for further research. Therefore, learner corpora could provide an unprecedentedly accurate description of the second language learners' data, which can help teachers discover more linguistic facts, provide feedback, and guide teaching.

Learner corpora could address linguistic issues of a particular learner population (Mukherjee & Rohrbach, 2006; Seidlhofer, 2002). Flowerdew (2001) promoted instructors to implement "insights gleaned from learner corpora… to complement those from expert corpora for syllabus and materials design" (p. 364), and Granger (2004) emphasized the necessity for more publications calling for the use of learner corpora to inform pedagogical practice. Thus far, with the development of more than one hundred learner corpora, the interest in using them has steadily increased (Cotos, 2014). Numerous learner corpus findings have emerged from contrastive analyses of lexical, grammatical, phraseological, pragmatic, and stylistic features of learner language (Granger, 1996). Previous research demonstrated that English language learners manifest problems with frequency, semantics, phraseology, register, and positioning (Gilquin et al., 2007). Although some of the findings are still at the level of implications (Granger, 2009), learner corpus insights are undoubtedly making their way toward successful learning and teaching.

As mentioned above, considering the importance of learner corpora for second language acquisition, second language teaching, and other fields, this paper aims to examine the deficiency and the potential of research on learner corpora. Put another way, with the help of the relevant citation index database, this study reviews the papers published in internationally renowned journals in the field of foreign learner corpus research from 2007 to 2021 and conducts a statistical analysis of research trends, research content, discipline, and language distribution, whose aim is to analyze and evaluate the development and impact of international research on learner corpora. The paper expects to inspire the development of and research on domestic learner corpora to improve the efficiency of foreign language teaching by evaluating and analyzing the research status and development trend of overseas learner corpora.

II. Research Design

A. Research Methods

This paper employed the WoS database as the source of literature, "learner corpus" and "learner corpora" as keywords, and the time from 2007 to 2021 as a period, initially 782 articles were searched. Following this, it used "Linguistics" further as the research orientation and category, and "Article" and "Review" as the document types, finally 615 related papers were retrieved as valid data for statistics and analysis. Then, literature was categorized and analyzed based on an overall trend, research content, subject, language distribution, and source journals to determine the research characteristics and prospects for future researchers at home and abroad.

B. Research Questions

Based on the statistical analysis of the retrieved literature, this study intends to answer the following questions:
(1) What is the general trend of overseas research on learner corpora in the past 15 years?
(2) Which countries and languages feature the most in overseas research on learner corpora?
(3) What is the main content of overseas research on learner corpora?

C. Research Tools: The Built-In Function of WoS for Analysis of Results

D. Results and Analysis

(a). The Overall Trend of Foreign Learner Corpora Research

Figure 1 shows the statistical results of the literature related to the research on learner corpora from 2007 to 2021

As seen in Figure 1, the research on learner corpora from 2007 to 2021 could divide into two stages.

The first stage refers to the period from 2007 to 2015. The articles published annually show an increasing trend from 11 in 2007 to 51 in 2015, peaking in 2016 with a total of 67 papers. There are three reasons for the increase in the overseas research on learner corpora: the development of corpora such as the large-scale web-based corpora, general-purpose corpora, and personalized, specialized, and industrialized small corpora (Li, 2010; Zhang & Chen, 2016); the diversified research on learner corpora, including psychology, neuropathology, literature, among others; and an increase in the number of journals that issue papers related to research on learner corpora. The top ten journals include System (39), CALL (33), Academic English (32), Language Learning (31), International Corpus Linguistics (28), Language Learning Technology (27), English for Special Purposes (25), Modern Languages (25), RECALL (23).

The second phase is from 2016 to 2021. The research shows a steady downward trend from 67 in 2016 to 53 articles in 2021. One explanation for this could be the difficulty in researching learners' language, such as the challenge of systematically testing it. Such reasons may affect research in this field. Researchers, teachers, and students should thus pay attention to this decline for two reasons. One reason is that learners' language and relevant research is one of the core topics in second language acquisition. The other is that learners' language features particularity and dynamics stemming from many factors such as the learner and their learning and social environments. In sum, there is a lack of

attention to learner corpora. However, such a situation also provides ample opportunities for further research in this field.

(b). The Distribution of Journals

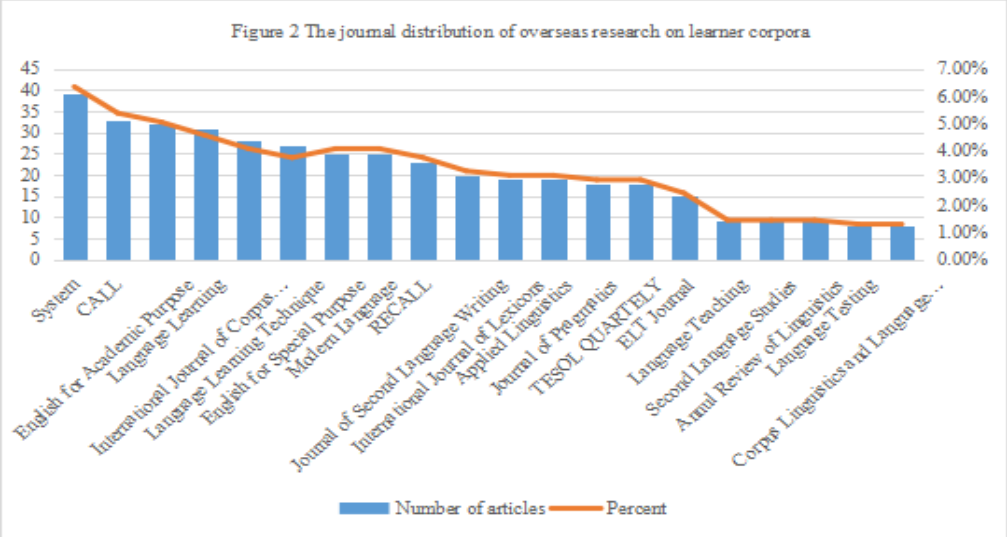

Figure 2 displays the distribution of journals for overseas research on learner corpora and their percentage from 2007-2021

As shown in Figure 2, overseas research on learner corpora appears wide in distribution and mainly in 20 internationally renowned journals. The top ten journals include System, CALL, English for Academic Purposes, Language Learning, International Journal of Corpus Linguistics, Language Learning Technology, English for Special Purposes, Modern Languages, and RECALL. However, journals like Second Language Writing, International Journal of Lexicography, Applied Linguistics, Language Teaching, Second Language Research, Language Teaching, and Language Testing published less. This distribution indicates that research on learner corpora needs to be concerned with relevant interdisciplinary research.

(c). Which Countries Feature the Most in Overseas Research on Learner Corpora

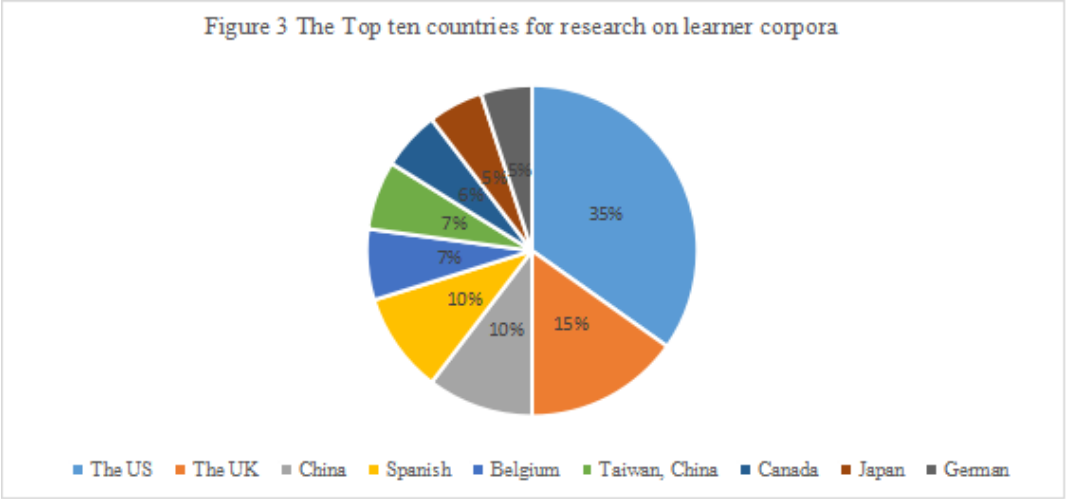

Figure 3 represents the top ten countries for research on learner corpora

Figure 3 shows that, among the top ten countries conducting and publishing research on learner corpora, there was an imbalance in the number of papers published, the US was the most prolific with 181 papers during the last 15 years, accounting for about 34.8% of all papers published in the same period. The country published approximately 2.3 times the number of the second-ranked UK, which published 79 articles (15.19%). Following these two countries, China issued 54 articles (10.38%), then Spain, Belgium, Taiwan, Canada, Japan, and Germany, publishing 51, 36, 35, 30, 28, and 26 papers, respectively. Worth mentioning is that although ranking third in the world, China had a considerable international influence on learner corpora research. Despite ranking third, it was still far behind the US and has only published twice as much as Germany which ranks last. The gap between China and other countries (or regions) was insignificant (within five percentage points). Therefore, the above evidence demonstrates that there was far enough

research on learner corpora given the number of multiple foreign language learners in China.

(d). Which Languages Feature the Most in Overseas Research on Learner Corpora

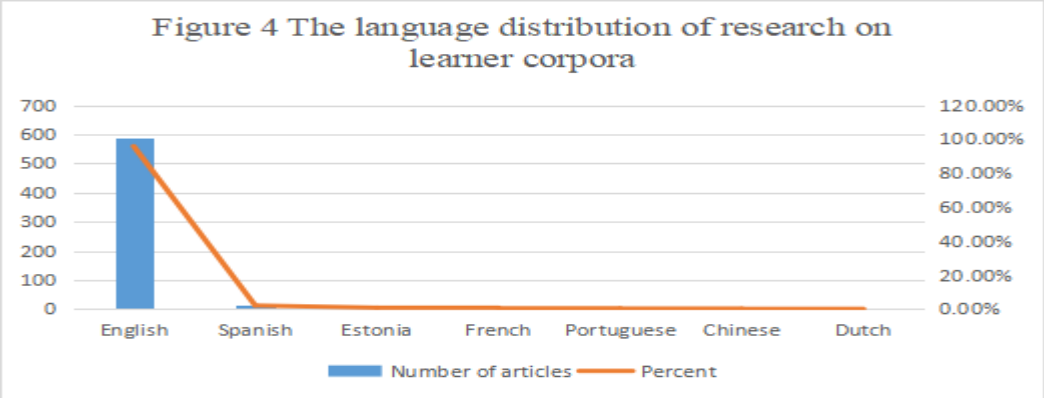

Figure 4 The language distribution of research on learner corpora

Known from the distribution of language shown in Figure 4, the overseas research on learner corpora was dominated by English, with 590 papers published in the language, accounting for 99.5% of the relevant research in the last 15 years. In contrast, fewer papers were published in other languages. Studies published in Spanish ranked second with only 13 articles, accounting for 0.22%, and those in other languages were even less. In particular, there was only one study based on Chinese learner corpora, accounting for only 0.17%. Under these circumstances, it is necessary to improve and strengthen the awareness of and ability to research learner corpora in China.

(e). The Discipline Distribution of Research on Learner Corpora

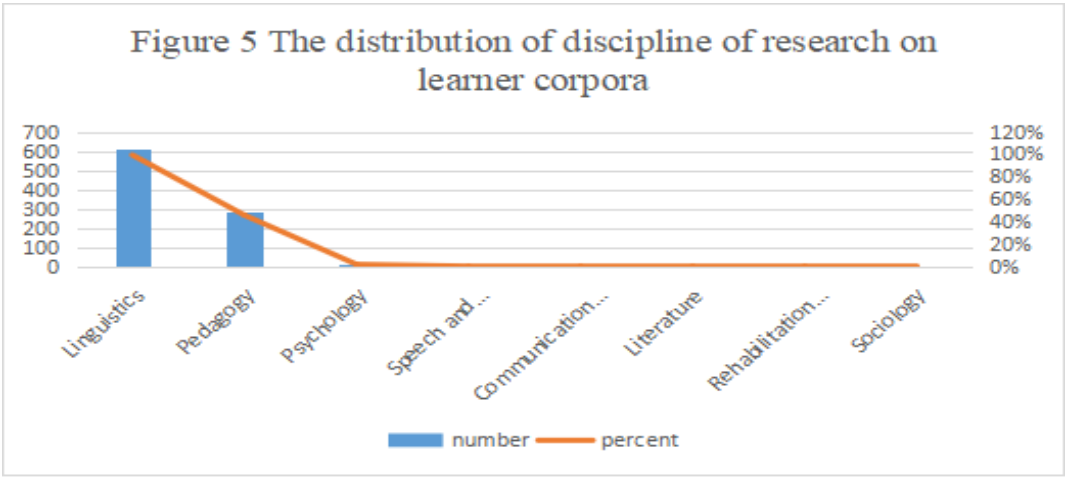

Figure 5 The distribution of discipline of research on learner corpora

As shown in Figure 5, the disciplines involved in research on learner corpora were not widely distributed, covering Linguistics, Education and Teaching, Psychology, Literature, and Sociology. Transparently, as one of the core topics in second language acquisition, research on learner corpora does not involve second language acquisition as an independent discipline. Of course, it might be categorized into linguistics research by default. However, when the research on learner corpora was extensive enough, the emerging discipline of second language acquisition might also be counted as an independent discipline by default in the statistics of the distribution of disciplines.

(f). The Content of Research on Learner Corpora

TABLE 1
THE CONTENT DISTRIBUTION OF OVERSEAS RESEARCH ON LEARNER CORPORA

Broad categories	Sub-categories
Introduction and Reviews (96 articles) (16%)	learner corpora (including introduction, building, and retrieval of and research on learner corpora, 59 papers, 9.7%)
	learner corpora and lexicography: (30 articles, 3.8%)
	learner corpora and translation: (7 articles, 1.7 %)
Research on Learner language (384 articles) (63%)	vocabulary: (including words, collocations, phrases, bundles, and the like): (157 articles, 4.1%)
	writing: (63 articles,1.6%)
	grammar: (53 articles,1.4%)
	speech: (46 articles, 1.2%)
	pragmatics: (18 articles, 0.5 %)
	testing: (17 articles, 0.4%)
	reading: (5 articles, 0.2%)
	listening: (3 articles, 0.1%)
	comprehensive category, including literature, style, textbooks, and phonetics, among others (22 articles, 0.6%)
Teaching (research)	82 papers (13%)
Linguistics (research)	53 articles (8%)

In terms of the research content of learner corpora, Table 1 shows that 96 (16%) papers were introductory, of which 59 (9.7%) are relevant to learner corpora (including part-of-speech tagging, syntax tagging, retrieval skills, etc.), 30 (3.8%) were dictionary-related, and 7, translation-related. In contrast, research on language accounts for the vast majority of research in this field, with a total of 384 papers published, accounting for 63%. Specifically, the language research mainly covers the four basic skills of foreign language learning (i.e., listening, speaking, reading, writing), vocabulary, grammar, and pragmatics. Additionally, learner corpora are used for research on teaching, with 82 related papers accounting for 13% of the total. The number of linguistics research is also considerable, with a total of 53 papers.

Overall, there are many studies on vocabulary acquisition based on learner corpora, less research on grammatical items and discourses, and almost no research on learners' phonology. In addition, although some studies have investigated the use of connectives in learners' compositions (e.g., Altenberg & Tapper, 1998; Pan & Feng, 2004), only a few have addressed thinking patterns and discourse structures (Deng, 2006). The vocabulary research only focused on the acquisition of words and some collocations, while that of multi-word units was not been widely studied. This is more or less consistent with Culpeper et al.'s (2018) findings that research on pragmatics based on learner corpora is heavily biased towards patterns that are immediately detectable in corpora. The LCR's Bibliographic Index currently lists 1144 studies based on learner corpora over the past 30 years, of which only 111 (9.7%) examined discourses or spoken languages, and only 16 focused primarily on pragmatics. By contrast, at least 85 (7.4%) papers focused on grammatical features, while 148 articles (12.9%) explored issues related to vocabulary or lexicography.

III. CONCLUSION

This study, employing the WoS database, conducted a statistical analysis of the status and trend of international research on learner corpora in the past 15 years. The results showed a mixed trend in international research on learner corpora from 2007 to 2021, with an increase in the first stage (2007-2015) and a slow decline in the second (2016-2021). The countries which featured the most in research on learner corpora were mainly Europe and the United States, among which the latter ranks first. In addition, previous research was dominated by English, with other languages being less involved, which is consistent with the general trend of research on the international corpus by Zhang et al. (2016). The journals publishing papers on learner corpora are widely distributed, mainly in 20 internationally renowned journals such as System, Journal of English for Academic Purpose, Language Learning, International Journal of Corpus Linguistics, Second Language Writing, Applied Linguistics, Language Teaching, and Second Language Research. Further, the research content is diverse, covering vocabulary, grammar, pragmatics, lexical chunks, teaching, translation, and others. However, most studies still focused on vocabulary acquisition, few studies on discourse pattern acquisition, and almost no studies on learners' phonology.

Despite the advantages of learner corpora and some findings achieved, we should never ignore their shortcomings and problems as manifested in the following six aspects: First, research on learner corpora has its problems. As pointed out by Wang et al. (2007), learner corpora could only provide static data in either written or spoken forms, they could not provide dynamic information about learning processes; they could only be employed to describe learners' ability to produce language (speaking and writing), but could not deeply examine learners' ability to comprehend language (listening and reading); learners' differences such as learning strategies and motivation could not be supported by learner corpora, either. This is consistent with Deng's (2007) study that, although learner corpora provide much evidence for language acquisition, it is only suitable for examining linguistic items present in corpora but not for those absent from there. Second, due to the scarcity of longitudinal learner corpora, there's no balance between sectional and longitudinal research. Most previous research is to discover the characteristics of learners' language at a particular stage

by comparing learners' corpora and L1 corpora. Studies on the development of learners' language based on longitudinal corpora need to be more focused. Third, most of the research on learner corpora is dominated by English, with other languages being less involved. As demonstrated above, many problems in research on learner corpora seem almost impossible to be solved, such as the development of learners' language proficiency (since there are no open longitudinal corpora available). Fourth, as mentioned above, research on learner corpora is not diverse enough, most of them are about introductions or reviews, but research on the development and construction of corpora is relatively insufficient, and research on some linguistic levels is even blank. Most papers with a high proportion published in this period are not diversified, focusing on language teaching. According to Sinclair (2004), studies on language based on learner corpora should focus more on spoken language. However, there's a small proportion of relevant studies published.

As this study has shown, learner corpora are a powerful resource for exploring learners' use of a foreign language. Over the last 15 years there has been an increasing focus on learner corpus research, with attention paid to how learner corpora could inform materials and approaches for L2 teaching (e.g., Granger et al., 2015). Although learner corpus research is still a very young field, more studies are called for to generate some interesting results. Learner corpora have become increasingly prominent in language learning and teaching, enhancing data-driven learning pedagogy, and learner corpus research has made remarkable development over the last 15 years, providing valuable scholarly insights (Charles, 2018). As Nesselhauf (2004) pointed out: "For language teaching... it is not only essential to know what native speakers typically say, but also what the typical difficulties of the learners of a certain language, or rather of certain groups of learners of this language, are" (p. 125). Learner corpora could help to reveal such difficulties and the differences between learners' language production and those features characterizing native-like language use.

Despite findings from learner corpus research, some areas are still largely underexplored. For example, there is a need for more investigations into learners' spoken language which is not only interesting in its own right but also allows for comparisons with learners' written language (Paquot & Granger, 2012). Additionally, more longitudinal studies are called for to identify features of learners' language development, being spoken or written language. Further, more studies following the lines of Nesselhauf (2009), Gilquin and Granger (2011), and Gotz and Schilk (2011) are needed to conduct comparisons of phrasicon in L1 and L2 varieties of English. A better knowledge of learner corpus research might profoundly benefit language teaching and learning since the field of learner corpus research is vast and has numerous potential applications. We hope to have demonstrated in this study that learner corpora provide a versatile resource to this end.

To sum up, this study is significant in understanding the overall trend of the development and characteristics, problems, and deficiencies of international research on learner corpora and in promoting relevant research.

REFERENCES

[1] Altenberg, B., & Tapper, M. (1998). The use of adverbial linking adverbials in advanced Swedish learners' written English. In S. Granger (ed.), *Learner English on Computer*. (pp. 80-93). Harlow: Addison Wesley Longman.

[2] Callies, M. (2015). Learner corpus methodology. In S. Granger, G. Gilquin, & F. Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*. (pp. 35-56). Cambridge, UK: Cambridge University Press.

[3] Charles, M. (2018). Book review. Learner Corpus Research: New Perspectives and Applications, Vaclav Brezina, Lynne Flowerdew (eds.). Bloomsbury Academic, London/New York. (pp. xvii+pp. 1-179). *Journal of English for Academic Purposes, (35)*, 1-3.

[4] Cotos, E. (2014). Enhancing writing pedagogy with learner corpus data. *ReCALL*, *26*(2), 202-224.

[5] Culpeper, J., Mackey, A., & Taguchi, N. (2018). *Second Language Pragmatics: From Theory to Research*. London, UK: Routledge.

[6] Deng, Y. C. (2006). A Study on the use of conjunctive adverbs in Chinese college students' English argumentative essays. *Teaching English in China*, *29*(6), pp. 32-36.

[7] Deng, Y. C. (2007). A review of learner corpus and second language acquisition research. *Foreign Language World*, *118*(1), pp.16-21.

[8] Flowerdew, L. (2001). The exploitation of small learner corpora in EAP materials design. In: Ghadessy, M. and Roseberry, R. (eds.) *Small Corpus Studies and ELT*. (pp. 363-379). Amsterdam: John Benjamins.

[9] Gilquin, G. et al. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, *6*(4), pp. 319-335.

[10] Gilquin, G., & Granger, S. (2011). From EFL to ESL: Evidence from the International Corpus of Learner English. In M. Hundt & D. Mukherjee (eds.), *Exploring Second-language Varieties of English and Learner Englishes: Bridging a Paradigm Gap* (pp. 55-78). Amsterdam, the Netherlands: John Benjamins.

[11] Gotz, S., & Schilk, M. (2011). Formulaic sequences in spoken ENL, ESL, and EFL. In M. Hundt, & J. Mukherjee (eds.), *Exploring Second-language Varieties of English and Learner Englishes: Bridging a paradigm gap* (pp. 79-100). Amsterdam, the Netherlands: John Benjamins.

[12] Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In Aijmer, K., Altenberg, B. and Johansson, M. (eds.) *Languages in Contrast. Text-based Cross-linguistic Studies*. (pp. 37-51). Lund: Lund University Press.

[13] Granger S. (2004). Computer learner corpus research: Current status and future prospects. In Connor U and TA Upton (eds.). *Applied Corpus Linguistics: A Multidimensional Perspective* [C]. (pp. 123-145). Amsterdam & Atlanta: Rodopi.

[14] Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In Aijmer, K. (ed.) *Corpora and Language Teaching*. (pp. 13-32). Amsterdam/Philadelphia: John Benjamins.

[15] Granger, S., Gilquin, G., & Meunier, F. (2015). Introduction: learner corpus research-past, present, and future. In S. Granger, G. Gilquin, & F. Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*. (pp. 2-5). Cambridge, UK: Cambridge University Press.
[16] Leech, G. N., & Garside, R. (1998). Running a grammar factory: The production of syntactically analyzed corpora or' treebanks'. Johansson, S. & Stenstrom, A. B. (ed.) *English Computer Corpora*. (pp. 29-31). Berlin/New York: Mouton de Gruyter.
[17] Li, W. Z. (1999). Corpus, learner corpus and foreign language teaching. *Foreign Language World*, *73*(1), pp. 51-55.
[18] Li, W. Z. (2010). A Critical Review of CIA. *Technology Enhanced Foreign Language Education, 127*(5), pp. 13-19.
[19] Mukherjee, J., & Rohrbach, J.-M. (2006). Rethinking applied corpus linguistics from a language pedagogical perspective: New departures in learner corpus research. In: Kettemann, B. & Marko, G. (eds.) *Planning, Gluing and Painting Corpora: Inside the Applied Corpus Linguist's Workshop*. Frankfurt: Peter Lang. 205-232.
[20] Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In: Sinclair, J. (ed.), *How to use Corpora in Language Teaching*. Amsterdam: John Benjamins. 125-152.
[21] Nesselhauf, N. (2009). Co-selection phenomena across new Englishes: Parallels (and differences) to foreign learner varieties. *English World-Wide*, *30*, 1-26.
[22] Pan, F., & Feng, Y. J. (2004). A corpus-based analysis of connectors in non-English Major Graduate Students' Writing. *Modern Foreign Languages*, *27*(2), pp. 157-162.
[23] Paquot, M., & Granger, S. (2012). Formulaic Language in Learner Corpora. *Annual Review of Applied Linguistics*, *32*, 130-149.
[24] Pravec, N. A. (2002). Survey of learner corpora. *ICAME*, *22*(26), pp. 81-114.
[25] Seidlhofer, B. (2002). Pedagogy and local learner corpora: Working with learning-driven data. In: Granger, S., Hung, J. and Petch-Tyson, S. (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. (pp. 213-234). Amsterdam/Philadelphia: Benjamins.
[26] Sinclair, J. (2004). *Trust the Text*. London: Routledge.
[27] Wang, L. F., & Zhang, Y. (2007). A corpus-based study of high-frequency verbs in Chinese learners' EFL writing. *Foreign Languages and Their Teaching*, *39*(2), pp. 110-116.
[28] Zhang, J. D., & Chen, W. (2016). A Visualized Analysis of International Corpus Linguistics Research. *Technology Enhanced Foreign Language Education*, *172*(12), pp. 66-73.

Shune Yang was born in Gansu, China in 1983. She is studying for her doctorate in The School of English Studies at Shanghai International Studies University, Shanghai, China.

Her research interests include Second Language Acquisition, Corpus Linguistics, English Teaching.