THE JOURNAL OF ASIA TEFL
Vol. 16, No. 2, Summer 2019, 718-726
http://dx.doi.org/10.18823/asiatefl.2019.16.2.20.718

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

Learner Corpora: Their Potentials for the Language Learning Classroom in Indonesian Primary School Contexts

Evynurul Laily Zen
Universitas Negeri Malang, Indonesia

Effendi Kadarisman
Universitas Negeri Malang, Indonesia

Aulia Apriana
Universitas Negeri Malang, Indonesia

Rahmati Putri Yaniafari
Universitas Negeri Malang, Indonesia

Introduction

Continuous investigations on language acquisition and development have been put in place to construct scientific explanations on how children comprehend language(s) of their surroundings as well as produce their own. With particular attention to studying multilingual children's language, we can gain knowledge on how multiple languages interact during the process of acquisition. Furthermore, in-depth examination toward typical errors in learners' L2/Ln, patterns of L1 influence, L1/L2/Ln vocabulary richness, and so forth may help us delve into the unique processes of being multilingual. In this context, our intuition suggests that by building a big data or so-called corpus on multilingual children's language production, it can assist any linguistic observations toward multilingual learning processes that are often demanding.

We cite McEnery and Wilson (1996) in defining corpus as a collection of written texts or transcribed speech. It is also important to consider Teubert and Cermakova's (2007) claim that corpus linguistics sees language as social phenomena. Thus, the act of collecting texts is to provide empirical evidence on natural language use for wider linguistic analyses (Kennedy, 1998). Keeping this notion in mind, our current research is projected not only to create big data on learners' language but also to conduct analysis on them for the particular purposes of capturing unique stages of development experienced by multilingual children in Indonesia to be able to contribute to language teaching and learning.

Referring to Timmis (2015), our *learner corpora* covers three major areas of development; (1) English learner corpora, (2) teaching-oriented corpora, and (3) the direct use of corpus data by learners or so-called Data Driven Learning. In this case, we extend our corpus into collecting the production of Indonesian and Javanese as the learners' first and second language. A significant impact of learner corpora in language pedagogy has been apparent to scholars as a relevant resource to evaluate existing

teaching material, a rich authentic linguistic resource for learners, and empirical evidence of learners' stages of language development (McEnery & Wilson, 1996; Timmis, 2015).

A number of scholars have put efforts into developing a corpus from various resources and for a variety of purposes: such as Tsay (2007) with the Taiwanese Child Language Corpus (TAICORP), Perera and Miranda (2010) with a voice corpus of children with language problems, Maryani (2011) with a corpus of Indonesian children's storybooks, and Prasad (2013) with a spoken lexical corpus of children in the Kannada language. Whenever children's language corpus is concerned, CHILDES is the most established data bank to date which has brought several influential benefits for child language studies (Macwhinney, 2000).

Considering the critical importance of learner corpora, we carried out a pilot project collecting written language production data from learners in two primary schools; (1) Primary Laboratory School of *Universitas Negeri Malang* and (2) *Surya Buana Malang*. Both schools are located in the urban area of *Malang* East Java. We believe that our corpus embodies not only a typical L1, L2, and L3 development of our learners but also a manifestation of the curriculum implemented in both schools. More importantly, our data bank will equip teachers with a well-documented learners' language that can help them look more closely at the point of departure as well as arrival achieved by their learners, so much so that teachers can make use of learner corpora to evaluate learners' acquisition progress.

Learner Corpora in Language Pedagogy

As a growing subject, corpus linguistics can be somewhat elusive from a well-framed definition. O'Keeffe and McCarthy (2010) maintain that corpus linguistics on the surface is associated with collections of texts, wordlists, and computer software. However, building a corpus itself is not an ultimate goal in linguistic research but a scientific tool to generate linguistic data for further analysis as it contains a relatively large data set (Kennedy, 1998).

In general, a corpus is created according to its purpose. Based on work by Nelson (2010) and Cheng (2012), large corpora are constructed upon need; either to conduct descriptive analysis on the whole language use or to explore a specific feature of the targeted language. In regards to corpus specifications, Sinclair (2004) elaborates that corpus has to be built by selecting a text based on its communicative purpose, including the entire text from the samples of language, and considering any detailed information of the text, such as parts of speech, typography, and layout.

Nelson (2010) states that some of the most influential corpus-based studies in the first half of the 20th century contain pedagogical purposes, such as those studies conducted by Thorndike (1921) on the creation of 4.5 million words of English for teaching literacy to native speakers of English in the United States. Adding to this major work, Timmis (2015) and Tono (2009) have listed a number of English-based corpora as follows;

a. JEFLL (Japanese EFL Learner): a corpus of 700.000 words collected from essays written by over 10.000 Japanese-speaking learners of English;
b. ICLE (International Corpus of Learner English): a multi-L1 corpus compiled by the University of Louvain from argumentative essays of the higher intermediate to advanced learners of English from several mother tongue backgrounds;
c. English Profile: a multi-L1 written and spoken corpus compiled by Cambridge University and collected from students all over the world;
d. LINDSEI (Louvain International Database of Spoken English Inter-language): a corpus of spoken English produced by advanced learners of English from several mother tongue backgrounds; and
e. TTT (Teachers Telling Tales): a corpus of teachers' languages in telling anecdotes task.

Learner corpora have been utilized extensively in pedagogical areas. Here are some previous works to date. Using corpus data, a pre-assumption on the possibility of teaching Key Stage 2 students English grammar and vocabulary has been drawn (Sealey & Thompson, 2004), a foreign use of English discourse markers by adult language learners has been identified (Friginal, Lee, Polat, & Roberson, 2017; Polat, 2011), a non-native use of English prepositions has been discovered through a comparative analysis of local corpora (Vienna corpus) and ICLE (Rankin & Schiftner, 2011), amongst others. These findings have served as a baseline for the exploitation of corpus data for instructional material design, syllabus/curriculum design, and language testing. In addition, a bridging line between corpus linguistics research and language pedagogy has been established (Cotos, 2014; Granger, 2003; Romer, 2011).

Method

In reference to Kennedy (1998), Francis (1992), and O'Keeffe and McCarthy (2010), we framed our research primarily within the area of corpus development and exploration. Thus, we employed a corpus-driven approach where we initially created our own corpus. We collected data by conducting an in-class writing task in three school subjects; Indonesian, English, and Javanese to all students of two private primary schools in *Malang* East Java.

To reduce the heterogeneity of the data, we selected a guided topic of "Tell me about your favorite toys" as the writing prompt. Our data analysis typically reflected a process of corpus building, such as transcribing learners' essays on Notepad with the format of .txt UTF8. It required us to type exactly as it was, together with errors, typos, symbols, and some other inaccuracies in order to keep the data original. This original record is essential when investigating children's language development. Following Reppen (2010) we designed our corpus by creating files at the smallest 'unit' meaning and each of the learner's essays was transcribed and stored as an individual file rather than as a whole class. While transcribing, we created metadata by including detailed information such as name, school, grade, gender, age, mother tongue, title of assignment, and homeroom instructor. Linguistic analyses were further carried out in AntConc as a corpus tool (Anthony, 2014).

Results and Discussion

As we collected texts in the languages spoken by bilingual children in Indonesia, we named our corpus "CBLING" which stands for a *Corpus of Bilingual Learners' Languages*. CBLING contains 64,426 word tokens in total, made up from a corpus on Indonesian language (505 essays and 23,724 tokens), Javanese language (531 essays and 24,159 tokens), and English language (497 essays and 16,543 words). Our corpus is similar to JEFLL in that it employs one writing topic for all participants as we considered it important to keep variables comparable (Tono, 2009). In addition, we also focused on collecting language data only from primary school students aged 6 to 12 years old with either Indonesian or Javanese as the L1 through in-class writing sessions.

Our learner corpora provide digitized language data that are useful for grammatical annotation analysis. This annotation will aid researchers in investigating the development of L1/L2/L3 grammars in multilingual children. In other words, analysing grammatical and lexico-grammatical patterns will be facilitated by the annotated corpora (Jones & Waller, 2015). Annotations, according to O'Keeffe & McCarthy (2010), may include tagging and annotating grammatical categories to create more effective and efficient linguistic analysis. To this extent, the annotation software is only available for English (Sagae, Davis, Lavie, & Macwhinney, 2010) and Indonesian (Suryawati, Munandar, Riswantini, Abka, & Andria, 2018) data sets. We therefore put the linguistic annotation of Javanese data into our shortlisted agenda. As such, our future research agenda will likely be similar to that of Sagae et al (2010) who have initially annotated the English section of the CHILDES database by referring to the grammatical relations

which provide insights into the automatic measurement of syntactic development in children. Their contribution to the annotated corpora has become very significant because it incorporates POS (Part of Speech) tags including ADJ (adjective), ADV (adverb), CO (communicator), CONJ (conjunction), DET (determiner), FIL (filler), and so forth. In short, the actual results of their project were the production of 37 grammatical relations that are crucial for the analysis of children's syntactic development, such as double-object construction, subject omission, verbal complements, quantifier scope, etc., which will equip the existing tools such as Developmental Sentence Scoring (Lee, 1974) or the Index of Productive Syntax (Sagae, Lavie, & Macwhinney, 2005; Scarborough, 1990)

Our learner corpora will make a significant contribution, particularly to the educational field, as Timmis (2015) has noticed a shift in function of a corpus in ELT from an unfamiliar to an inseparable part of English language teachers' daily needs. It is because when the corpus-based research is introduced to teachers, they will have an effective toolkit to investigate as well as predict their students' language development. Tejada, Gallardo, Ferrada, and Lopez (2015), for example, developed an automatic language proficiency assessment using CLEC (CEFR-Labeled English Corpus) collected from the CEFR-based levels of English texts produced by L2 learners of English. Most teachers may also need to look at estimated numbers of words their students have to acquire, relevant grammatical aspects their students need to master, relevant phrases their students need to practice by conforming to any corpus available online or their own corpus compiled from their students' language performances.

Following the elaborate aims of the CEP Corpus and the JLE Corpus (Timmis, 2015), our learner corpora can also be utilized to identify the following components: (a) typical English L3-specific errors, (b) overuse or underuse of syntactic features compared to adult English, (c) tendency of cross-linguistic influence between languages, (d) different levels of L1/L2/L3 proficiency, and (e) different stages of L1/L2/L3 development. We believe our learner corpora can be a pathway to explore many aspects of language use across different linguistic levels. Moreover with Uria, Maritxalar and Zabala's (2014) discovery on a computer application to analyze language errors, teachers and language learners find it beneficial for the teaching/learning process. In regards to typical grammatical errors, using Anthony's (2014) AntConc software (Figure 1) demonstrates a preliminary finding on the production of auxiliary word *am* in CBLING where we can exploit these findings to evaluate learners' progress. The concordance analysis can also inform us of patterns of auxiliary *am* across different grades that are beneficial for materials evaluation and teaching strategies.

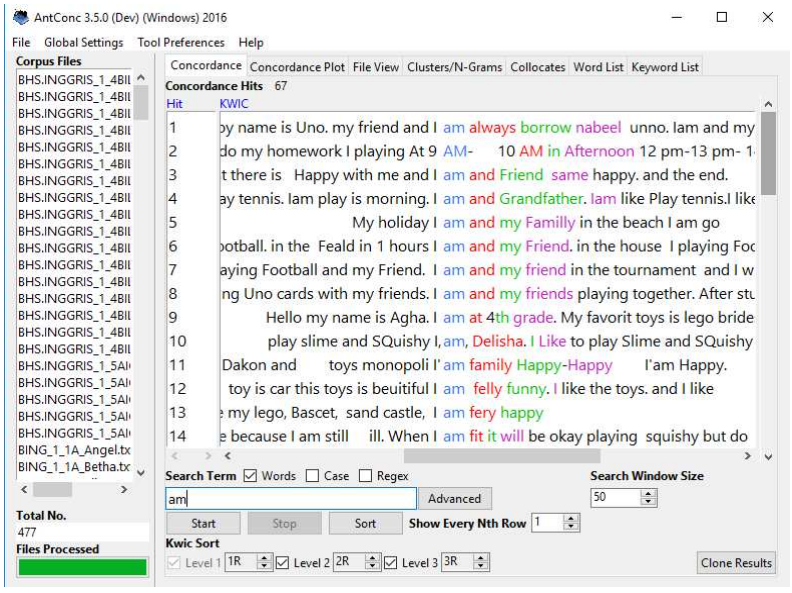

Figure 1. Excerpt of typical production of English auxiliary.

Attempting to find some insights on the utilization of our corpus, we refer to Laakso and Smith (2007) who worked on the acquisition of verb meanings using corpus data to examine the regularities of words and word relations. They observed the natural use of pronouns in relation to the meaning of the verb that follows by considering that pronouns are the most frequently used closed-class words in spoken English as both syntactic subjects and objects in both adult and children's speech. In other words, by discovering patterns of pronouns, their corpus-based research seeks to confirm patterns of verb meanings. Segbers and Schroeder (2016), on the other hand, utilized corpus data to investigate vocabulary development that is crucial as it may reflect on other aspects of language such as grammar (Bates & Goodman, 1999) and phonology (Gathercole & Baddeley, 1989) to reading comprehension (Ouellette, 2006; Tannenbaum, Torgesen, & Wagner, 2006) and reading ability and school success (Muter, Hulme, Snowling, & Stevenson, 2004). Corpus-based vocabulary research conducted by Kotani, Yoshimi, and Isahara (2013) was also influential as it created sets of bilingual word cards in 9 different language combinations. Making sense of these previous works, we believe that CBLING has the potential to assist such an investigation of L1/L2/L3 vocabulary complexity as well as measure a weaker and stronger language as our multilingual participants acquire Indonesian, Javanese, and English most probably in different acquisition circumstances. Figure 2 illustrates sentence production patterns of Javanese where teachers can carry out a careful observation on the development of Javanese vocabularies and the possibility of language mixing.

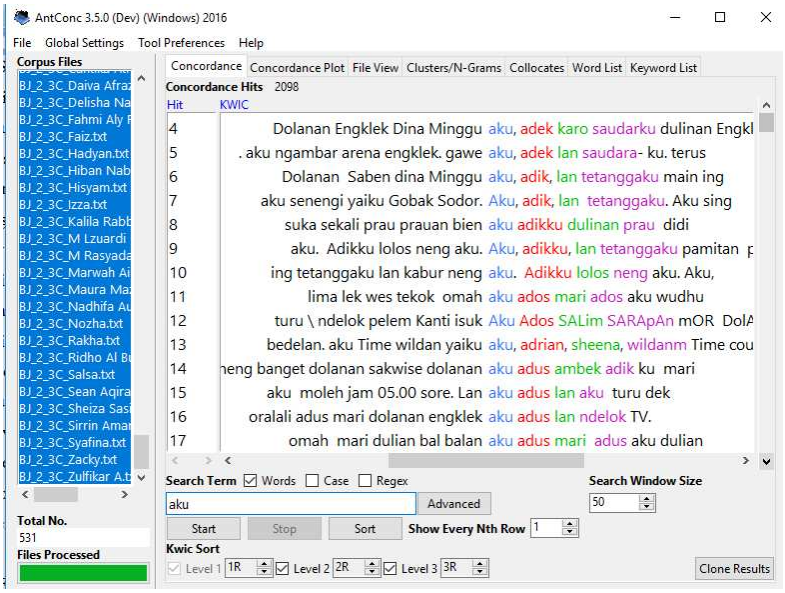

Figure 2. Excerpt of Javanese language essays.

Segbers and Schroeder (2016) studied the relationship between test performance and lexicon size using a language corpus in German. Their study suggested an important finding on vocabulary development during primary school and maintained the benefits of corpus data for a comparative analysis of vocabulary size in children and adults where children's vocabulary size was seen to contain 55% nouns, 35% verbs and 10% adjectives while adults' vocabularies were about 60% nouns, 30% verbs and 10% adjectives. This comparison might inform us on the growth of vocabulary across one's life span. More importantly, by referring to this corpus finding, teachers could gain knowledge on their learners' vocabularies in order for them to develop far richer and more relevant learning materials as well as create a more engaging learning experience.

Bringing corpus data into the classroom, Maryani (2011) developed a core vocabulary of English for pre-school students by firstly building a limited corpus of Indonesian children's storybooks and examining the most frequent nouns, verbs, adjectives, and adverbs to finally be translated into a number

of relevant English words for Indonesian preschoolers. We, on the other hand, built our own learner corpus to be exploited for learners and by learners themselves.

Beyond all possibilities that our corpus could provide, we believe that the key strength of CBLING is that it keeps records of naturally occurring language data even though it was produced with limitations as we controlled the topics and genres of the written language. In addition, our corpus presents words in their context, not in isolation. Words of a language in corpus data are basically sourced from natural language use in the speakers' daily communication, not merely as a list. This claim has been made very clear by Cheng (2012, p. 151) who argues that "Corpus linguistics provides lots of evidence for the ways in which words have distinct and describable patterns regarding the company they keep, and the company they do not keep".

Conclusion

Based on the results of our research, we conclude that we were successful in putting the project in place. The CBLING (Corpus of Bilingual Learners' Languages) is a collection of in-class writing tasks conducted in three languages; Indonesian, English, and Javanese. Compiled from multilingual learners in Indonesia enrolled in Grades 1 – 5 of two private primary schools in the urban area of *Malang*, East Java, we have a relatively large corpus of 64,426 tokens in total from 1,533 essays.

As it is the first corpus containing multilingual learners' natural language production, CBLING has become a pioneer in child language data bank in Indonesia and will benefit primary school teachers, practitioners, and language enthusiasts in particular, in examining learners' language acquisition and development to further be exploited using any linguistic frameworks for several pedagogical purposes such as curriculum design and evaluation, materials development, assessment, and so forth.

Acknowledgement

We thank the anonymous reviewers for their fruitful feedback.

The Authors

Evynurul Laily Zen (corresponding author) is currently a PhD candidate at the National University of Singapore and also an academic staff at Universitas Negeri Malang, Indonesia. Her research projects mainly include topics on multilingualism and multilingual education.

Department of English
Universitas Negeri Malang
Malang, Jalan Semarang No. 5, Indonesia
Tel: +62 341 551312
Mobile: + 62 85648489390
Email: evynurul.laily.fs@um.ac.id

A. Effendi Kadarisman earned his Ph.D degree in linguistics at the University of Hawaii in 1999. His research interests include (a) themes of universality and relativity in linguistics, (b) applied linguistics, and (c) poetics and ethnopoetics. He is currently a professor of linguistics at Universitas Negeri Malang, Indonesia.

Department of English
Universitas Negeri Malang

Malang, Jalan Semarang No. 5, Indonesia
Tel: +62 341 551312
Mobile: + 62 81331452486
Email: achmad.effendi.fs@um.ac.id

Aulia Apriana is currently a lecturer in the Department of English at Universitas Negeri Malang, Indonesia. Her interests cover micro linguistics (Phonology and Syntax) and some aspects of macro linguistics (Sociolinguistics and Pragmatics).

Department of English
Universitas Negeri Malang
Malang, Jalan Semarang No. 5, Indonesia
Tel: +62 341 551312
Mobile: + 62 8155506613
Email: aulia.apriana.fs@um.ac.id

Rahmati Putri Yaniafari is an academic staff member at the Department of English, Universitas Negeri Malang. She has a keen interest in English Language Teaching, especially Autonomous Learning, Computer Assisted Language Learning (CALL) and Content and Language Integrated Learning.

Department of English
Universitas Negeri Malang
Malang, Jalan Semarang No. 5, Indonesia
Tel: +62 341 551312
Mobile: + 62 81253950672
Email: yaniafari.fs@um.ac.id

References

Anthony, L. (2014). *AntConc* (Version 3.4.3) (Computer Software). Tokyo: Waseda University.

Bates, E., & Goodman, J. . (1999). On the emergence of grammar from the lexicon. In B. Macwhinney (Ed.), *The emergence of language* (pp. 29-80). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Cheng, W. (2012). *Exploring corpus linguistics: Language in action*. London and New York: Routledge.

Cotos, E. (2014). Enhancing writing pedagogy with learner corpus data. *ReCALL*, *26*(2), 202–224. http://doi.org/10.1017/S0958344014000019

Francis, W. (1992). Language corpora. In J. Svartvik (Ed.), *Trends in linguistics: Studies and monographs 65* (pp. 17–32). Berlin & New York: Mouton de Gruyter.

Friginal, E., Lee, J. J., Polat, B., & Roberson, A. (2017). *Exploring spoken English learner language using corpora: Learner talk*. Palgrave Macmilan.

Gathercole, S., & Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language*, *28*, 200–213.

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.

Jones, C., & Waller, D. (2015). *Corpus linguistics for grammar: A guide for research*. London and New York: Routledge.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.

Kotani, K., Yoshimi, T., & Isahara, H. (2013). Application of reading data in an integrated learner corpus. *Procedia - Social and Behavioral Sciences*, *95*, 513–521. http://doi.org/10.1016/j.sbspro.2013.

10.676

Laakso, A., & Smith, L. B. (2007). Pronouns and verbs in adult speech to children: A corpus analysis. *Journal of Child Language*, *34*, 725–763. http://doi.org/10.1017/S0305000907008136

Lee, L. (1974). *Developmental sentence analysis*. Evanston, IL: Northwestern University Press.

Macwhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, New Jersey: Lawrence Erlabaum Associates.

Maryani. (2011). Identifying Indonesian-core vocabulary for teaching English to Indonesian preschool children : A corpus-based research. *K@ta*, *13*, 147–162.

McEnery, T., & Wilson, A. (1996). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Muter, V., Hulme, C., Snowling, M. J., & Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: Evidence from a longitudinal study. *Developmental Psychology*, *40*(5), 665–681. http://doi.org/10.1037/0012-1649.40.5.665

Nelson, M. (2010). Building a written corpus: What are the basics? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 53-65). New York: Routledge.

O'Keeffe, A., & McCarthy, M. (Ed.). (2010). *The Routledge handbook of corpus linguistics*. London and New York: Routledge.

Ouellette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology*, *98*(3), 554–566. http://doi.org/10.1037/0022-0663.98.3.554

Perera, G., & Miranda, C. (2010). Voice corpus in Spanish for children with language problems. In *Proceedings of the Ninth Mexican International Conference on Artificial Intelligent*, 137-142. http://doi.org/10.1109/MICAI.2010.26

Polat, B. (2011). Investigating acquisition of discourse markers through a developmental learner corpus. *Journal of Pragmatics*, *43*(15), 3745–3756. http://doi.org/10.1016/j.pragma.2011.09.009

Prasad, B. A. M. (2013). Use of markers observed in the spoken language lexical corpora of children in Kannada language. *Language in India*, *13*(7), 456–473. Retrieved from www.languageinindia.com

Rankin, T., & Schiftner, B. (2011). Marginal prepositions in learner English: Applying local corpus data. *International Journal of Corpus Linguistics*, *16*(3), 412–434. http://doi.org/10.1075/ijcl.16.3.07ran

Reppen, R. (2010). Building a corpus: What are the key considerations? In M. O'Keeffe, & A. McCarthy (Ed.), *The Routledge handbook of corpus linguistics* (pp. 31-37). London and New York: Routledge.

Romer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics*, *31*, 205–225. http://doi.org/10.1017/S0267190511000055

Sagae, K., Davis, E., Lavie, A., Macwhinney, B., & Wintner, S. (2010). Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, *37*, 705–729. http://doi.org/10.1017/S0305000909990407

Sagae, K., Lavie, A., & Macwhinney, B. (2005). Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (pp. 197–204). Ann Arbor, MI: Association for Computational Linguistics.

Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics*, *11*(1), 1–22.

Sealey, A., & Thompson, P. (2004). 'What do you call the dull words ?' Primary school children using corpus-based approaches to learn about language. *English in Education*, *38*(1), 80–91.

Segbers, J., & Schroeder, S. (2016). How many words do children know ? A corpus-based estimation of children' s total vocabulary size. *Language Testing*, 1–24. http://doi.org/10.1177/0265532216641152

Sinclair, J. (2004). *Trust the text: Language, corpus, and discourse*. London and New York: Routledge.

Suryawati, E., Munandar, D., Riswantini, D., Abka, A. F., & Andria, A. (2018). POS-tagging for informal language (study in Indonesian tweets). In *International Conference on Data and Information*

Science. IOP Publishing Ltd.

Tannenbaum, K. R., Torgesen, J. K., & Wagner, R. K. (2006). Relationships between word knowledge and reading comprehension in third-grade children. *Scientific Studies of Reading*, *10*(4), 381–398. http://doi.org/10.1207/s1532799xssr1004

Tejada, M. A. Z., Gallardo, C. N., Ferrada, M. C. M., & Lopez, M. I. C. (2015). Building a corpus of 2L English for automatic assessment: The CLEC Corpus. In *7th International Conference on corpus linguistics: Current work in corpus linguistics: Working with traditionally-conceived corpora and beyond (CILC 2015)* (pp. 515–525). Elsevier B.V. http://doi.org/10.1016/j.sbspro.2015.07.474

Teubert, W., & Cermakova, A. (2007). *Corpus linguistics: A short introduction*. London: Continuum International Publishing Group.

Thorndike, E. L. (1921). *The teacher's word book*. New York: Columbia University Press.

Timmis, I. (2015). *Corpus Linguistics for ELT: Research and practice*. New York: Routledge.

Tono, Y. (2009). Integrating learner corpus analysis into a probabilistic model of second language acquisition. In P. Baker (Ed.), *Contemporary corpus linguistics* (pp. 184-203). London: Continuum International Publishing Group.

Tsay, J. S. (2007). Construction and automatization of a Minnan Child Speech Corpus with some research findings. *Computational Linguistics and Chinese Language Processing*, *12*(4), 411–442.

Uria, L., Maritxalar, M., & Zabala, I. (2014). An environment for learner corpus research and error analysis: The study of determiner errors in Basque. *International Journal of Computer-Assisted Language Learning and Teaching*, *4*(3), 34–51. http://doi.org/10.4018/ijcallt.2014070103