Exploring the Grammatical Development of Multilingual Learners of English: A Corpus-Based Perspective

Khilda Husnia Abidah, Elisa Ratih, Evynurul Laily Zen(✉),
Ira Maria Fran Lumbanbatu, and Anisatul Ilmiah

Universitas Negeri Malang, Malang, Indonesia
evynurul.laily.fs@um.ac.id

Abstract. Our current study looks carefully at the development of grammatical learning of English among multilingual young learners in Indonesia. Our specific objectives include examining typical grammatical errors in the learners' writing across different genres in the corpus, and observing the production of sentential negation construction in the English datasets. The study is set out to conduct a corpus exploitation with the learning corpora itself – we use CBLING (Corpus of Bilingual Learners' Languages), in this case – containing 154.496 word-tokens from around 1,016 English short essays (Zen et al., 2017). Our initial findings indicate that grammatical errors typically appear around the production of tense markers, person features, and negation. Taken together, our findings are essential not only to inform teachers of English on the learners' stages of language development, but also for them to design relevant pedagogical interventions. The CBLING itself has become a pioneer in child language data banks that will benefit primary school teachers and language acquisition enthusiasts for further linguistic and pedagogical exploration.

Keywords: Learner corpora · CBLING · Grammatical development

1 Introduction

Our present study departs from a unique yet complex language development of multilingual learners in Indonesian context, with a focus on English. Among hundreds of local languages and a national language, English has received a special place in the multilingual ecology of this country due to the fast growth of communication and technology. Especially in large urban centres, according to Kealing and Wiradisastra [1], the engagement of English had engaged in the media and youth repertoires resulting in a significant number of English-based nativization such as *mbois* 'boyish' and *nyentrik* 'eccentric' [2], lexical borrowings such as *bisnis* 'business', *fiks* 'fixed', *kredit* 'credit' [3], and various forms of language alternation practices such as code-mixing and switching [4]. The very warm welcome to English is also due to the institutionalization of this language as the sole official working language of ASEAN [5], facilitating the massive utilization

U. Widiati et al. (Eds.): ASIATEFL 2022, ASSEHR 749, pp. 311–320, 2023.
https://doi.org/10.2991/978-2-38476-054-1_26

of the language in a few important elements in ASEAN societies' life including education and public communication. By considering the exponential growth of English, the question we propose prior to the implementation of this current research was how English takes shape in multilingual children repertoire in Indonesia, assuming that they have been receiving exposure to English from classrooms, the Internet, and media. It is therefore essential to take steps forward to document the development or the acquisition outcomes of their English which, in practice, becomes very challenging to do with respect to the individual variations each child might have.

The documentation of a large number of children language production has been initiated by CHILDES with this language database growing to be the most established data bank to date and serving as an important source of language studies [6]. CHIDLES has been used as a resource for several studies since its first establishment in 1984 by Brian WacWhinney and Catherine Snow. Extensive studies include the production of temporal adverbs among English-speaking children [7], word and sound errors in young speakers [8], the complex verbs in German, Dutch, and English [9], the acquisition of English questions [10], the acquisition of English dative construction [11], the acquisition of English adjective lexicon [12], modality, infinitives, and finite bare verbs in Dutch and English [13], and the finiteness systems and lexical aspects in child Polish and English, and hundred others [14].

Since the corpus building and corpus exploration started to be on the rise, this scientific tradition has been extended to educational contexts with the establishment of learner corpora – a specific type of corpus that collects learners' language(s). A significant number of learner corpora have been built for various purposes, such as the Arabic Learner Corpus [15], the Barcelona English Language Corpus [16], a longitudinal trilingual corpus of young learners of Italian, German and English or LEONIDE [17], and many more highly influential English-based learner corpora being discussed in the following section.

Given the increasing demand of learner corpora, *CBLING* (Corpus of Bilingual Learners' Languages) of Universitas Negeri Malang (UM) has been developed to collect spoken and written language production of multilingual young learners in Indonesia [18]. CBLING has also facilitated several pedagogical investigations, for example, a pilot study on pronominal use and tense production for the teaching of Javanese and English [19], typical errors in learners' second language production and possible patterns of cross-linguistic influence [18], and an exploration of learners' narrative ability as reflected in the macrostructure element of their writings [20]. Our current project extends the previous works on CBLING by focusing on the grammatical development of these young multilingual learners.

Here, we project the use of learner corpora as one of the ways to make this analysis more effective. As a collection of learners' natural language use, learner corpora have facilitated teachers in assessing the development of narrative skills among multilingual learners is often demanding due to individual variations and various other factors. In this context, the use of big data or so-called corpora is highly potential in assisting teachers and researchers in mapping out learners' narrative development. As a collection of written texts or transcribed speech collected from learners' natural language use, learner corpora have served as an important data source for wider linguistic analyses [21, 22] and established a linkage between theory and practice primarily in language teaching and learning [23, 24, 19].

Monolingual and bilingual children differ with respect to language development, with the second group showing a general tendency of lag in the age of acquisition of some aspects in the target language and cross-linguistic influence from other languages [25]. Regarding the second trait, at the syntax level, transfer is broadly described as any instances of influence of a syntactic structure of a language in a syntactic production and/or perception of another language in bi/multilingual speakers.

By referring to CBLING as the primary data source and emphasizing in the English datasets of multilingual learners, we aim to explore their grammatical development with a focus on examining tense production. Following Lucero [26], by putting this project in place, we intend to articulate a potential significance of learner corpora for pedagogical purposes. Here, our findings will inform English teachers of several important insights especially learners' grammatical development as reflected in their narratives which can be used to design relevant pedagogical interventions. In addition, we framed our research primarily within the area of corpus development and exploration that the results will provide theoretical contribution in the understudied areas of corpus analysis [22, 27, 28].

Specifically, our study is set out to examine typical grammatical errors in the learners' written and spoken production with a focus on past tense production, and how the corpus data can inform the potential similarities and differences.

2 Method

We adopted a corpus-based approach in which we carried out a linguistic exploration on a readily available corpus data. In this case, we use CBLING (Corpus of Bilingual Learners' Languages) that has been developed by the Research Group on Linguistics (RoLING) at the Department of English, Universitas Negeri Malang (UM) these past five years.

CBLING contains 154.496 word-tokens sourced from a collection of 1.016 essays written by multilingual learners in their background languages: Indonesian, Javanese, and English [18]. More than 500 learners from seven different primary schools in East Java – SD Laboratorium UM Malang, SD Laboratorium UM Blitar, MI Al-Akbar Surabaya, SD Muhammadiyah Manyar Gresik, SD Muhammadiyah Ikrom Wage Sidoarjo, SD Laboratorium UNESA Surabaya, and SDI Surya Buana Malang participated in the project. In addition to the written datasets, CBLING also collected the spoken data in which the elicitation process was conducted through several experimental tasks that include (1) spoken picture naming, (2) spoken story production, (3) spoken storytelling, (4) written gap filling, (5) written story retelling, and (6) written storytelling in Javanese and English. For the purpose of the current analysis, however, we focused on analysing the corpus of spoken story production and written story retelling as our study aims to map out typical errors on past tense with a brief comparative analysis between spoken and written language production. Both datasets were in English.

We utilized AntConc – a free corpus tool – to elicit the target production from CBLING with two grammatical pointers being the focus of elicitation. They were verbs with past markers in each of the datasets – spoken and written, and inflected verbs with agreement markers indicating person features. In addition, our corpus analysis was also set out to observe verb frequency in both datasets. These analyses were drawn to inform readers the potential differences and similarities of learners' spoken and written production.

3 Findings and Discussion

The empirical findings on past tense and person-feature errors in our corpus analysis will be presented in two categories based on the nature of speech production: written and spoken datasets. In each of the datasets, we will provide a summary table demonstrating the overall production as well as the percentage of accurate (target-like) and inaccurate (non-target like) production in two subcategories: regular and irregular verbs. Following the table, samples of production will be presented and discussed within the framework of relevant literature.

3.1 Evidence of Verb Tense Errors in the Written Corpus

In the written corpus, the results indicate that past verbs were produced 1.341 times with the regular verbs (189 tokens) tending to be much smaller in frequency than the irregular one (1152 tokens) (see Table 1). Table 1 also informs us that, despite the big number of production (1.341 tokens), it can be seen that there was not much variation as only 6 types of different regular verbs and 10 types of irregular verbs appeared in the corpus. That means that, on average, one regular verb was used 31 times, while, on the other hand, one irregular verb was in use about 115 times.

With respect to the percentage of past tense accuracy, we learn that non-target-like or inaccurate production was greater than the target-like ones for both regular and irregular verbs. More specifically, 85% of inaccurate use is significantly larger than 13% of accurate use of regular past verbs. Interestingly, it appears that the comparative percentage in irregular past verb production is much smaller: 64% of errors and 33% of accuracy were shown.

Our corpus findings seem to provide us a hint on the nature of morphological operation by young learners, especially in the acquisition of regular-irregular verbs. In the case of English monolingual children, they generally produce bare verbs for all events at the initial stage of tense acquisition [29]. Their acquisition begins with the use of verbs

Table 1. Verbs with past markers in the **written** datasets

Past Verbs	Target-like	Non-target like
Regular Verbs (6 types, 189 tokens)	25 (13%)	164 (85%)
Irregular Verbs (10 types, 1152 tokens)	384 (33%)	768 (64%)

without grammatical markers. Then, at the age of two, these monolingual children start to use morphological markers, sometimes for tense and agreement features [30, 31]. In this stage, previous evidence shows that children tend to acquire irregular past verbs before the regular ones with some overregularization of irregular verbs appearing along the process that lasted around six to seven years of age [32]. While there may be an argument that monolingual and bilingual children should differ in the way they acquire grammatical patterns of a target language, extensive studies have also demonstrated that bilingual children follow the same acquisition pattern as monolingual peers with the bilingual groups lagging behind [25].

Some of the production samples indicating incorrect use of past verbs in the written corpus are the following.

(1) Then Susi and dad **sat** down and than dad **say** I **wanted** go home
(2) Silvian and father **go** home. He **said** happy birthday and they happy.

Data (1), interestingly, shows that the first and the third verbs – *sat* and *wanted,* respectively – were correct, yet the second one (*say*) was not. Data (2) appears to be similar in which the first verb (*go*) was incorrect while the second one (*said*) was surprisingly correct. It is important to underline that such productions are typical in the written corpus we investigated, meaning that the acquisition of verbs in their past forms among these young learners remain incomplete or on-going. Some of the past verb forms might successfully be acquired, while some others might not yet.

Regarding the significant number of errors made, following Andersen [33] and Andersen and Shirai [34] Aspect Hypothesis (AH), we argue that the early stage of grammatical learning is generally constrained by semantic aspect and not the grammatical morphology. It denotes that children rely heavily on lexical items to mark grammatical pointers rather than utilizing inflected verbs or attaching inflectional morphemes on a verb. Moreover, the typical errors we found in the datasets have been very predictive as in Indonesian, verbs are not inflected for any grammatical markers. As such, EFL learners seem to highly likely be influenced by the grammatical pattern of their background language, or Indonesian in particular.

With regard to the frequency level, Table 2 shows the five most frequent verbs in the written corpora. Two of them are regular verbs: *ask* and *want*, whereas the other three are irregular: *say, go,* and *buy.* In terms of production accuracy, the regular verb *want* (95%) was mostly inaccurate, followed by the irregular verbs *go* (94%) and *buy* (94%). The irregular verb *say* is interestingly found to be mostly accurate.

The variation we see among these five most frequent verbs can be interpreted from the perspective of token frequency effect. It explains that the frequency of use of a token – or also so-called lexical item – in the input of the target language determines the acquisition of this token. It implies that the more the lexical item appears in the speech of adults or any other forms of language input, the easier the learners to acquire such item. The greater appearance of a token leads to memorization [25]. In this case, the irregular verb *say* was indeed spoken highly frequently by the storyteller in the stimulus video we used for this experimental task. Therefore, we assume that learners received an adequate amount of accurate forms of *say* before they wrote their own.

3.2 Evidence of Verb Tense Errors in the Spoken Corpus

Further empirical evidence is from the spoken datasets in which the results can be seen in Table 3. The total past verbs produced was 2.518 verbs, which is interestingly almost twice larger than the written counterpart. It comprises 887 regular verbs and 1.631 irregular verbs. Regarding the variability, there were 24 different regular verb forms and 25 different irregular ones. On average, one regular verb is likely to appear 37 times and 65 times for each irregular verb.

As indicated in Table 3, both types of verbs are similar in terms of the fact that non-target-like production is higher than target-like ones. Interestingly, not only that the irregular verbs were found to be used more significantly, but also that they were less likely to be incorrect when compared to the regular verb production. To put it differently, it seems that young learners tend to produce more accurate irregular past verbs. In this case, our findings provide support for previous research, especially Marcus, Pinker, Ullman, Hollander, Rose, and Xu [32], highlighting the pattern of verbal acquisition in which irregular past verbs were generally acquired before the regular ones.

In terms of errors in past verb production, the percentages appear to be very high in both regular and irregular verbs, 84% and 68% respectively (see Table 3). Here, we argue that errors in the early stage of additional language learning might not only be interpreted from the fact that their acquisition process may still be in progress, but also that these young learners' background language(s) may influence the developmental process. As Gass and Selinker [35] maintain, most multilingual speakers often find it difficult to keep the knowledge and uses of their languages apart. Given that multiple linguistic resources interact during acquisition and development [36, 37], we assume that when learning English as a foreign language, children's previously learned languages – Indonesia and/or

Table 2. Most frequent past verbs in the **written** corpora

Past Verbs	Target-like	Non-target like
Say	329 (82%)	70 (18%)
Ask	18 (24%)	57 (76%)
Go	18 (6%)	295 (94%)
Want	5 (5%)	93 (95%)
Buy	9 (6%)	145 (94%)

Table 3. Verbs with past markers in the **spoken** datasets

Past Verbs	Target-like	Non-target like
Regular Verbs (24 types, 887 tokens)	144 (16%)	743 (84%)
Irregular Verbs (25 types, 1631 tokens)	517 (32%)	1114 (68%)

Table 4. Most frequent past verbs in the **spoken** corpora

Past Verbs	Target-like	Non-target like
See	41 (18%)	193 (82%)
Fall	179 (59%)	122 (41%)
Say	31 (24%)	100 (76%)
Look	14 (5%)	244 (95%)
Find	147 (48%)	162 (52%)

any local language – provide significant effects, especially when the grammatical patterns of these languages are different. In specific, while past tense markers are inflected to the verbs in English, these markers are totally isolated from the verbs in Indonesian. As a result, learners may find it challenging to move from one grammatical pattern to another at the beginning of their language learning.

In addition to the percentages of verb errors, we also attempted to look at the verb frequency from our spoken corpora. Table 4 demonstrates the top five verbs appearing in the corpus: one regular and four irregular verbs. Interestingly, the regular verb *look* (95%) appears to be the most significantly inaccurate. On the irregular verb part, the percentage of inaccuracy of *fall* (41%) tends to be the smallest one while *see* (82%) is the highest one.

Typical past verb errors we found in our bilingual children's speech production provide support for the role of cross-linguistic transfer as evident in several previous studies. For example, in their comparative analysis on the production of English past tense between Chinese–English and French–English bilingual children, found that the types of errors the children in both groups made were in the area of morphophonological level attributed to the knowledge of the other language the children had acquired before [25].

Taken together, our analysis indicated that the past verb errors in bilingual children's written and spoken production in CBLING are not significantly different. Irregular past verbs were produced more frequently than the regular ones in both corpora. Similar trends for both corpora were also seen from the fact that regular verb errors were more typically than the irregular verbs. A clear contrast between the two corpora is that the spoken corpus (2518 tokens) contains a greater number of word tokens than in the written one (1341 tokens). This empirical finding has lent a pedagogical significance, for example, as a reference in developing a more relevant learning media and material to facilitate the acquisition and development of L2 grammar. In this case, we follow the argument of Ayoun and Salaberry [38] that grammatical features are one of the key markers in children's syntactic and semantic development.

4 Conclusion

To conclude, our corpus analysis indicated that, in terms of the total number of tokens, the spoken corpus contains more past verbs than the written one. We also found that irregular verbs were used highly more frequently than the regular one in both written and spoken corpora. Interestingly enough, however, inaccurate production was highly likely in regular past verbs (84%–85%) than the irregular ones (64%–68%) in both types of datasets.

Our findings can be used to carry out need analysis prior to the development of a lesson plan and design a more relevant pedagogical intervention.

References

1. J. Kealing, and G. Wiradisastra, “English for Internationalization”, *ESL Magazine*. pp. 22–26, 2009. Available: www.eslmag.com.
2. S. Kartomihardjo, “Ethnography of communication codes in East Java”, *Pacific Linguistics*. vol. D(39), pp. 1–39,1979. https://doi.org/10.15144/PL-D39.cover.
3. P. H. Lowenberg, “English as an additional language in Indonesia”, *World Englishes*. vol. 10(2), pp. 127–138, 1991. https://doi.org/10.1111/j.1467-971X.1991.tb00146.x.
4. M. Da Silva, “Upon the prevalence of English on billboard advertisements: Analyzing the role of English in Indonesian contexts”, *TEFLIN Journal*. vol. 25(1), pp. 33–61, 2014.
5. Kirkpatrick, “English in ASEAN: Implications for regional multilingualism”, *Journal of Multilingual and Multicultural Development*. vol. 33(4), pp. 331–344, 2012. https://doi.org/10.1080/01434632.2012.661433.
6. MacWhinney, “The CHILDES project: Tools for analyzing talk (3rd ed.)”, *Mahwah*. New Jersey: Lawrence Erlabaum Associates, 2000.
7. M. Chiang, “The acquisition of three temporal adverbs by two English-speaking children”, *Journal of National Chengchi University*. vol. 60, pp. 19–56, 1989.
8. F. Wijnen, “Incidental word and sound errors in young speakers”, *Journal of Memory and Language*. vol. 31(6), pp. 734–755, 1992.
9. H. Behrens, “How difficult are complex verbs? Evidence from German, Dutch and English”, *Linguistics*. vol. 36, pp. 679–712, 1998.
10. E. Dabrowska, “From formula to schema: The acquisition of English questions”, *Cognitive Linguistics*. vol. 11, pp. 83–102, 2000.
11. L. Campbell, and M. Tomasello, “The acquisition of English dative constructions”, *Applied Psycholinguistics*. vol. 22(02), pp. 253–267, 2001.
12. Blackwell, “Acquiring the English adjective lexicon: relationships with input properties and adjectival semantic typology”, *Journal of Child Language*. vol. 32(03), pp. 535–562, 2005.
13. E. Blom, “Modality, Infinitives, and Finite Bare Verbs in Dutch and English Child Language”, *Language Acquisition*. vol. 14(1), pp. 75, 2007.
14. R. M. Weist, A. Pawlak, and K. Hoffman, “Finiteness systems and lexical aspect in child Polish and English”, *Linguistics*. vol. 47(6), pp. 1321–1350, 2009.
15. Alfaifi, and A. Atwell. (2012). *Arabic Learner Corpus* [Online]. Available: https://www.arabiclearnercorpus.com/
16. Muñoz, (ed.) *Age and the Rate of Foreign Language Learning*. Clevedon: Multilingual Matters, 2006.
17. Glaznieks, J.-C. Frey, M. Stopfner, L. Zanasi, and L. Nicolas, “LEONIDE: A longitudinal trilingual corpus of young learners of Italian, German and English”, *International Journal of Learner Corpus Research*. vol. 8(1), pp. 97–120, 2022.
18. E. L. Zen, A. Apriana, E. Kadarisman, and R. P. Yaniafari, “Learner corpora: Their potentials for the language learning classroom in Indonesian primary school contexts”, *The Journal of Asia TEFL. vol. 16*(2), pp. 718–726, 2019. DOI: https://doi.org/10.18823/asiatefl.2019.16.2.20.718

19. E. L. Zen, and A. Nurisnaini, “Exploiting third language production corpora for pedagogical purposes,” *Indexed E-proceedings of International Seminar on Language, Literature, and Culture,* pp. 1–19, March 2019. DOI https://doi.org/10.18502/kss.v3i10.3882.
20. E. L. Zen, “A corpus-based analysis on multilingual children’s narratives in Indonesian contexts”, *Lingua.* vol. 15(1), pp. 91–98, 2020. https://doi.org/10.18860/ling.v15i1.7731
21. T. McEnery, and A. Wilson, *Corpus linguistics.* Edinburgh: Edinburgh University Press, 1996.
22. G. Kennedy, *An introduction to corpus linguistics.* London: Longman, 1998.
23. S. Granger, “The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research”, *TESOL Quarterly.* vol. 37(3), pp. 538–546, 2003.
24. Timmis, *Corpus Linguistics for ELT: Research and practice.* New York: Routledge, 2015.
25. E. Nicoladis, S. Pika, and P. Marentette, “Do French-English bilingual children gesture more than monolingual children?”, *Journal of Psycholinguistic Research.* vol. 38(6), pp. 573–585, 2009. https://doi.org/10.1007/s10936-009-9121-7.
26. Lucero, “The development of bilingual narrative retelling among Spanish–English dual language learners over two years”, *Language, Speech, and Hearing Services in Schools.* vol. 49(3), pp. 607–621, 2018. https://doi.org/10.1044/2018_lshss-17-0152.
27. W. Francis, “Language corpora. In J. Svartvik (Ed.)”, *Trends in linguistics: Studies and monographs 65.* Berlin & New York: Mouton de Gruyter, 1992, pp. 17–32.
28. M. Nelson, “Building a written corpus: What are the basics? In A. O’Keeffe & M. McCarthy (Eds.)”, *The Routledge handbook of corpus linguistics.* New York: Routledge, 2010, pp. 53–65.
29. V. A. Marchman, “Children’s productivity in the English past tense: The role of frequency, phonology and neighborhood structure”, *Cognitive Science.* vol. 21, pp. 283–304, 1997.
30. R. Brown, *A first language.* Cambridge, MA: Harvard University Press, 1973.
31. Philips, “Syntax at age two: Some cross-linguistic differences,” *Papers on Language Processing and Acquisition. MIT Working Papers in Linguistics.* vol. 26, pp. 325–382, 1995.
32. G. F. Marcus, S. Pinker, M. T. Ullman, M. Hollander, T. Rose, and F. Xu, “Overregularization in language acquisition”, *Monographs of the Society for Research in Child Development.* vol. 57(4), 1992.
33. R. Andersen, “Developmental sequences: The emergence of aspect marking in second language acquisition. In T. Huebner & C. Ferguson (Eds.)”, *Crosscurrents in second language acquisition and linguistic theories.* Amsterdam: Benjamins, 1991, pp. 305–324.
34. R. Andersen, and Y. Shirai, “Discourse motivations for some cognitive acquisition principles”, *Studies in Second Language Acquisition.* vol. 16, pp. 135–156, 1994.
35. S. M. Gass, and L. Selinker, *Second language acquisition: An introductory course* (Third). New York and London: Routledge, 2008.
36. J. Rothman, “L3 syntactic transfer selectivity and typological determinacy: The typological primacy model”, *Second Language Research.* vol. 27(1), pp. 107–127, 2010. https://doi.org/10.1177/0267658310386439.
37. M. Sung, “Englishization in Asia: Language and Culture Issues, *World.* vol. 30(1), pp. 151–163, 2011.
38. Ayoun, and M. R. Salaberry, “Acquisition of English tense-aspect morphology by advanced French instructed learners”, *Language Learning.* vol. 58(3), pp. 555–595, 2008. https://doi.org/10.1111/j.1467-9922.2008.00450.x
39. O’Keeffe, and M. McCarthy, (Ed.) *The Routledge handbook of corpus linguistics.* London and New York: Routledge, 2010.