Conference Paper

Exploiting Third Language Production Corpora for Pedagogical Purposes

Evynurul Laily Zen[1] **and Alvi Nurisnaini**[2]

[1]State University of Malang, Indonesia
[2]Primary Laboratory School of State University of Malang, Indonesia

Abstract

Investigating bilingual learners' learning third language (L3) can be a daunting task for teachers, in the extent of individual differences. However, by documenting learners' L3 production in a corpus file format, certain analysis can be imparted more easily to examine every possible aspect that is at play during learning process. More specifically, when having a big size of natural speech data, teachers will have loads of empirical evidence of their learners' language to conduct a variety of scientific exploration on stages of language development, the argument we borrow from O'Keeffe, McCarthy and Carter (2007). A corpus of this kind is particularly useful for teachers in developing teaching-oriented corpora and for learners in having a direct contact to the corpus data or so-called 'data-driven learning' (Timmis, 2015). Our current paper, therefore, focuses on the exploitation of a learner corpus in the teaching and learning of third language in Indonesian context. Our corpus was built within the seven months of the main author's dissertation fieldwork conducted to the 261 students of grade 3 of six primary schools in East Java that has enriched the previous limited corpus of CBLING (Corpus of Bilingual Language). Collected through a variety of experimental tasks, these corpora compiled the English and Javanese L3 written production. The findings suggest that our learner corpora can be exploited for pedagogical purposes such as to provide learners with primary linguistic resources and authentic materials, to supply teachers with empirical evidence of common language errors and interlanguage performance as to enable them to monitor learners' L3 acquisition and development, to help teachers construct a more relevant lesson plan, to evaluate existing teaching materials, and so forth. In this way, we aim at promoting an innovative teaching and learning through 'big data' exploitation. In the realm of individual differences, the investigation of bilingual learners learning the third language (L3) can be a daunting task for teachers. However, by documenting their L3 production, an analysis that examines every possible aspect that is at play during the learning process can be conducted more easily. More specifically, when having a big size of natural speech data, teachers will have bundles of evidence of their learners' acquisition necessary to conduct scientific exploration on stages of language development. This is the argument borrowed from McCarthy and O'Keeffe (2010) who refer to the extensive use of CHILDES Language Database as first language research resources dating back as early as the 1960s. A corpus of this kind is especially useful for teachers in developing teaching-oriented corpora and for learners in having a direct contact with the corpus data or so-called 'data-driven learning'. In this study, our corpus was built during a seven-month dissertation fieldwork involving 261 students of Grade

Corresponding Author:
Evynurul Laily Zen
evynurul.laily.fs@um.ac.id

Received: 23 January 2019
Accepted: 26 February 2019
Published: 17 March 2019

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the International Seminar on Language, Education, and Culture Conference Committee.

How to cite this article: Evynurul Laily Zen and Alvi Nurisnaini, (2019), "Exploiting Third Language Production Corpora for Pedagogical Purposes" in *International Seminar on Language, Education, and Culture*, KnE Social Sciences, pages 1–19. DOI 10.18502/kss.v3i10.3882

3 of six Primary Schools in East Java. It compiled the English and Javanese L3 written production of all six schools and the English and Javanese L3 spoken production of two schools collected through a variety of experimental tasks. The findings suggest that learner corpora can be exploited for pedagogical practices such as to provide learners with primary linguistic resources and authentic materials, to supply teachers with empirical evidence of common language errors and interlanguage performance as to enable them to monitor learners' L3 acquisition and development, and to help teachers construct a more relevant lesson plan. This way, we aim to not only promote an innovative teaching and learning through a 'big data' exploitation but also elevate the interface of research and practice.

Keywords: learner corpora, L3 production, pedagogical practices

1. Introduction

Among a strong tradition of longitudinal and case-specific research, studies in language acquisition have moved further towards the incorporation of big data or so-called corpus into analysis. Within this seemingly increasing demand, corpus analysis, however, never ignores the essence of natural language data as it itself is a collection of natural language use. In this way, corpus analysis and language acquisition research have imposed a similar scientific belief in that both put and will always put natural language data as the best source of linguistic evidence (Sinclair, 1991). More importantly, in the specific context of learner corpus, such analysis is also seen to able to bridge the gap between corpus linguistics and second/foreign language acquisition research. Furthermore, it underlines the contribution it has made in linking its theoretical to practical implication mainly in the area of language teaching and learning (Granger, 2003; Timmis, 2015). To this direction, we have put our current research in place, where we want to see how a learner corpus we have formerly built can enhance third language learning. More explicitly, we aim at exploiting natural language data we had collected from classroom to be pedagogically useful for classroom.

We refer to Sundh (2016) who has pinpointed a sense of practicability in the exploitation of multilingual learner corpora in her scientific observation of cross-linguistic phenomena. From such corpora, vocabulary choices and linguistic features as a result from multidirectional interaction among learners' languages can be very valuable not only in understanding learner language development but also in imposing pedagogical insinuation (Shirato & Stapleton, 2007).

Several studies have addressed issues of using corpora in the analysis of multilingual learners' language. Polat (2011) examined the use of three focal discourse markers by an adult language learner over the course of one year naturalistic data collecting. From this developmental learner corpus, they figured out very different patterns of use of the markers. Focusing on different linguistic features, Rankin and Schiftner (2011) discovered a foreign use of English preposition from a comparative analysis of local corpora (The Vienna Database of English Learner Texts) to the International Corpus of Learner English. In Indonesian context, Maryani (2011) has initiated a corpus building of Indonesian children's storybooks as an effort of establishing Indonesian-core vocabulary for teaching English to Indonesian preschoolers. Her corpus was compiled from 131 Indonesian children's storybooks resulting in a corpus of 134,320 words which then be analyzed using MonoConc Pro to find frequent nouns, verbs, adjectives, and adverbs. Her findings have obviously sounded the importance of corpus in the teaching of English to preschoolers especially in a construction of a more appropriate instructional material.

Together with Romer (2011), we believe that the abovementioned findings have become a fundamental baseline for language pedagogy as well as given theoretical contribution to language acquisition. According to Granath (2018), the connecting link between corpus research and language teaching/learning was just initiated at the second Teaching and Language Corpora (TaLC) conference in Lancaster in 1996 in which one of the highlights was that corpus data can be exploited for instructional material design, in addition to syllabus/curriculum design, language testing, and classroom methodology (Granger, 2003; Cotos, 2014). The idea of incorporating corpus for teaching purposes is basically rooted from the fact that corpus evidence can suggest "which language items and processes are most likely to be encountered by language users, and which therefore may deserve more investment of time in instruction" (Kennedy, 1998, p. 281)

However, in spite of many publications as listed by Meunier and Littre (2013) that include Burnard and McEnery (2000), Sinclair (2004) and Connor and Upton (2004) on *the use of corpora in language teaching and learning*; Botley, McEnery and Wilson (2000) on *the use of multilingual corpora in teaching and research*; Granger, Hung and Tyson (2002) on *the links between computer learner corpora, second language acquisition and foreign language teaching*; Mukherjee and Rohrbach (2006) and O'Keeffe, McCarthy, and Carter (2007) on *the use of native and learner corpora in the classroom*, Römer (2018) himself was still in doubt whether corpora and corpus tools have been extensively implemented in teaching activities and/or whether teachers and learners are knowledgeable enough to use them. He carried out a survey asking teachers about resources they most likely used in exam marking. The result in Figure 1 below confirms that corpus is the least resource being employed by them.

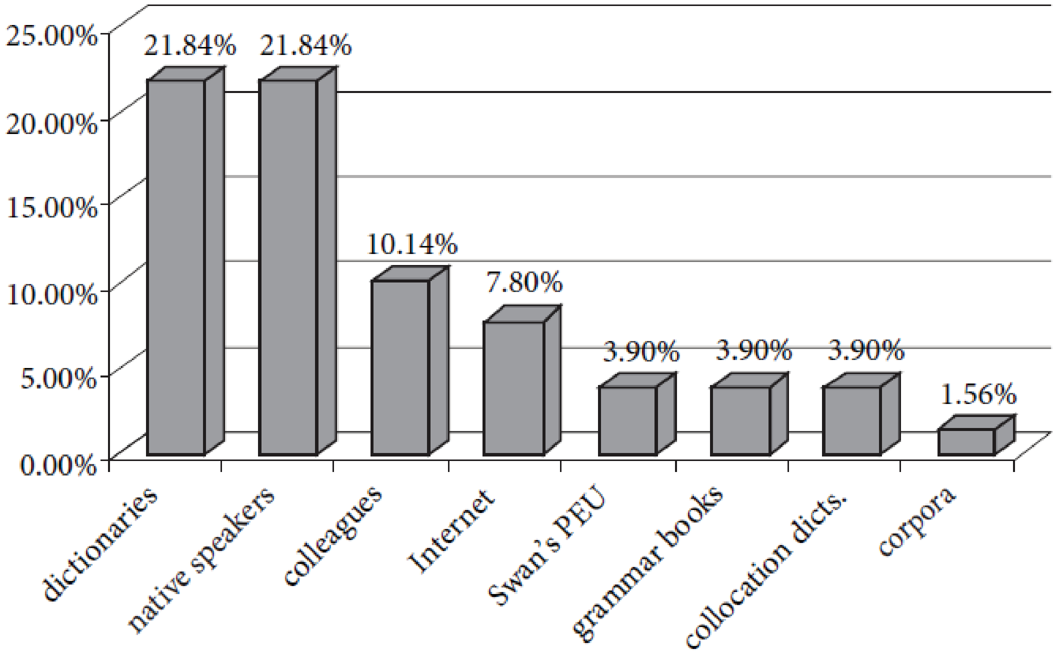


Figure 1: Resources consulted by teachers in exam markings (Romer, 2018).

Along these lines, our current paper is situated within the exploitation of a learner corpus in the teaching and learning of third language in Indonesian context. Our learner corpus was built from the language productions of multilingual learners speaking Indonesian, Javanese, and English. Some of them are considered to acquire English as an L3 with L1 Indonesian and L3 Javanese, while some others are seen to have Indonesian as an L1, English as an L2, and Javanese as an L3. Considering English and/or Javanese as the learners' L3, we will explore any possible linguistic evidence that may arrive from our learner corpora and how teachers can manipulate these evidences as teaching resources to mainly enhance L3 learning and acquisition.

2. Framework

We frame our paper solely on the nature of learner corpora in language teaching/ learning outlining brief definitions, types/examples, and pedagogical applications.

Corpus is generally defined as a collection of written text or transcribed speech (Kennedy, 1998; McEnery & Wilson, 1996). In the specific context of learner corpora, we refer to Granger (2003) who defines it as "electronic collections of authentic texts produced by foreign or second language learners". This specific type of corpus is commonly built with the aims of evaluating existing materials, providing learners with authentic linguistic resources, analyzing learners' languages, and measuring learners' language development (Timmis, 2015) which can be explored by multiple parties including researchers, authors, experts, teachers and students (Hana, Rosen, Stindlova, & Stepanek, 2014).

The first learner corpora were created in 1990s through the ICLE (International Corpus of Learner English) project. The project has embarked two major corpora in the field; the Longman Learner Corpus and the Cambridge Learner Corpus (Timmis, 2015). Containing 10 million words, the Longman Learner Corpus was collected from texts written by learners of English from different levels of proficiency and from twenty different L1 backgrounds. The texts include in-class essays, timed examination papers and other types of written assignment. The Cambridge Learner Corpus, on the other hand, was a collection of exam papers of learners of English taking Cambridge ESOL English examinations all over the world. It contains over 25 million words covering over 85 000 scripts from 180 countries and 100 different L1 backgrounds. These learner corpora have, to a large extent, facilitated language researchers as well as teachers with precious information about learners‘ mistakes, learners' interlanguage, typical L1-specific errors, overuse and underuse items, learners' different proficiency levels, etc (Ibid).

There have also been an increasing number of non native corpora across countries. We provide brief reviews of some of them here.

1. COLSEC (The College Learners' Spoken English Corpus in China)

   It is the first spoken English corpus of non-English major university students in China collected from the transcriptions of the College English Test-Spoken English Test (CET-SET) from 2000 to 2004, with a total of 723,299 tokens (Yang & Wei, 2005).

2. The first learner corpus of Czech

   The corpus was compiled from texts written by students of Czech as a second or foreign language and by near native young speakers of Czech with Romani background. The written parts of CzeSL and ROMi have now reached 2.2 million word tokens. It is in addition to short essays written by non-native learners of Czech and students with Romani background with 1.2 and 0.5 million tokens and theses written in Czech by foreign students with 0.5 million words (Hana et al., 2014).

3. CALES (Corpus Archive of Learner English in Sabah-Sarawak)

It contains 480,000 words of argumentative essays collected from university undergraduates studying in four institutions in the East Malaysia states of Sarawak and Sabah (UiTM's Sarawak and Sabah campuses, Universiti Malaysia Sarawak and Universiti Malaysia Sabah). Essays were written in class under timed conditions. Each was completed with a Learner Profile instrument providing personal, pedagogical and sociolinguistic information about the students (Botley, 2014).

4. The JEFLL Corpus

   It is a collection of free compositions written by more than 10,000 Japanese-speaking learners of English. The corpus size is approximately 700,000 words. It consists of the subjects ranging from novice to intermediate levels, covering mainly junior and senior high school students in Japan (Timmis, 2015).

In regards to the use of corpora for textbook, ELT publishers have a strong preference to use native corpora for the reason that they contain real and authentic English (Meunier & Gouverneur, 2018). Macmillan, for example, considers using World English Corpus because it is a unique modern database of over 200 million words revealing fresh information on how words are used and natural examples of English as it is written and spoken now (Ibid). However, we agree to Cotos (2014) who maintained that this type of corpora should not be used for syllabus design as they cannot indicate what L2 features are difficult and problematic for learners. For this particular purpose, non native speaker corpora will serve as a more appropriate resource as it contains authentic texts produced by L2 learners which therefore can be used to reveal learners' difficulties, individual differences and specific characteristics of non native use (Ibid). It is arguably important to also refer to Nesselhauf (2004: 125) who pointed out that "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."

Apart from the development of both native and non native learner corpora, the regular use of these big data especially in the EFL classroom is still uncommon. It is because learning how to use corpora is rarely, not to mention never, a part of teacher training courses (Granath, 2018). It, thus far, becomes crucial to explicitly impart the pedagogical use of learner corpora in which we will refer to Romer (2011) below.

In Figure 2, he explains two possible applications of corpus in language teaching and learning; direct and indirect. The direct application, also termed as *data-driven learning* (DDL), is where teachers and students can actively access corpus data as well as use corpus tools by themselves to assist the teaching and learning process. In DDL,

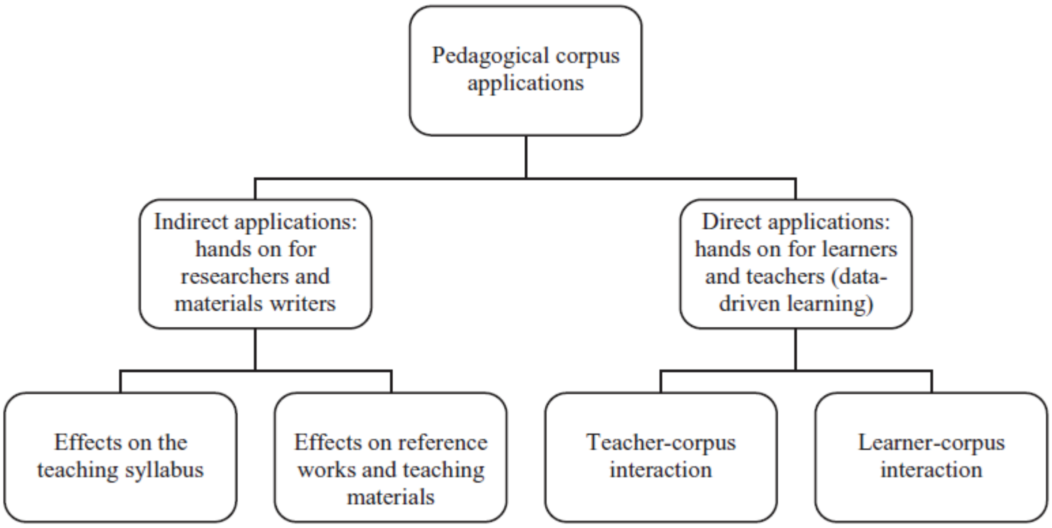

Pedagogical corpus applications
Indirect applications: hands on for researchers and materials writers
Direct applications: hands on for learners and teachers (data-driven learning)
Effects on the teaching syllabus
Effects on reference works and teaching materials
Teacher-corpus interaction
Learner-corpus interaction

Figure 2: Pedagogical corpus applications.

according to Cotos (2014), students can directly analyze corpus data in the form of concordances, frequency lists, keyword lists, clusters, etc., for them to extract patterns of L2/Ln use. Furthermore, Tim Johns, who pioneered direct corpus applications in grammar and vocabulary classes at the University of Birmingham (UK) in the 1980s, suggested to “confront the learner as directly as possible with the data, and to make the learner a linguistic researcher” (2002, p. 108). Borrowing Johns’ term, Römer (2018) referred the learners in this learning situation as a “language detective”. Learner corpora can also be indirectly applied and have an effect on the teaching preparation. In this context, teachers working on course, syllabus, and material design can refer to results of previous corpus analyses. In other words, while the direct approach is more teacher/learner-focused, the indirect approach centers on the impact of corpus evidence on the teaching and learning design.

In reference to how teacher can utilize corpora in their classrooms, we refer our current study partly to the following two previous studies. Granath (2018) has elaborated her experience of teaching EFL syntax at the university level using corpora particularly as an aid to design exercises and to make grammar more authentic in the classroom. She pinpointed that students are allowed to encounter ‘real language’ rather than made-up examples and teachers can employ them in different ways, such as to create exercises, demonstrate variation in grammar, discuss near-synonyms and collocations, etc. (Ibid). Botley (2014) analyzed critical thinking and argumentation abilities of students writing in a foreign language using CALES learner corpus (Corpus Archive of Leaner English in Sabah/Sarawak). They seek evidence on how Malaysian undergraduates structure their written arguments and found out the emerging issues of L1 interference particularly in the construction of grammar including the misuse of articles, the verb to be, and subject-verb agreement.

3. Method

We employed corpus-based approach where linguistic evidences generated from our corpus would be employed for teaching purposes. With the utilization of AntConc to locate keyness/keyword lists, clusters, and concordance analysis, we demonstrated how to use these corpus evidences in language classroom.

As aforementioned, our main corpus was compiled from the main author's dissertation fieldwork conducted to the 261 students of Grade 3 of six Primary Schools in East Java. The participants were particularly enrolled in (1) UM Lab School of Malang, (2) UM Lab School of Blitar, (3) MI Masjid Al-Akbar Surabaya, (4) SD Muhammadiyah "Ikrom" Wage Sidoarjo, (5) SD Muhammadiyah Manyar Gresik, and (6) UNESA Lab School. In specific, they were in International Class Program (ICP) of an SBI-type of school affiliated to the Cambridge Assessment International Education where both national and international curricula are integrated into their teaching/learning.

Collecting both written and spoken data, we carried out several experimental tasks that include (1) picture naming, (2) story production, (3) spoken storytelling, (4) gap filling, (5) story retelling, and (6) written storytelling in both English and Javanese. At the moment of writing this paper, our corpus has reached 154.496 tokens collected from 1.016 essays. This corpus enriches the previous small-scale corpus of CBLING (Corpus of Bilingual Language) resulted from a pilot project funded by the State University of Malang, Indonesia in 2017. With the total of 20.251 tokens, CBLING was compiled from short written essays in Indonesian, Javanese, and English language composed by the students of grades one to five in two public primary schools in Malang, Indonesia; *Surya Buana Private Primary School* and *UM Lab School* (Apriana, Kadarisman, & Yaniafari, 2017).

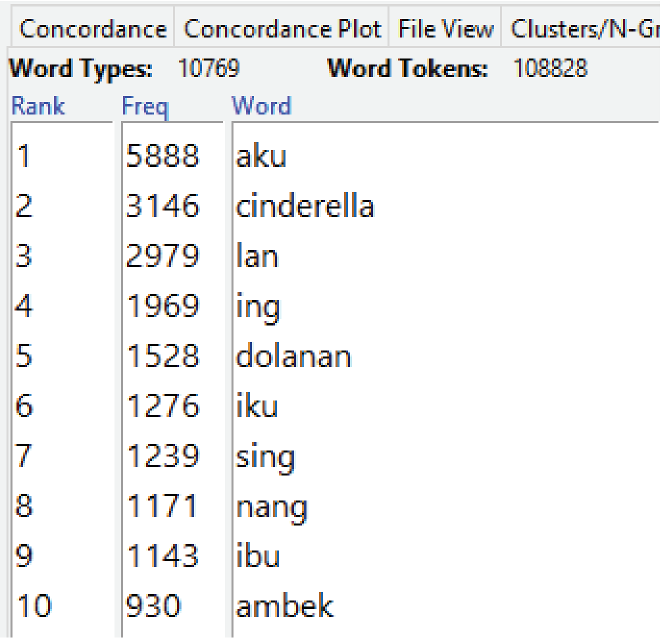
In regards to the learner profiles, the data we collected data from Language Background Questionnaires (LBQ) indicated that most learners come from middle to upper class families assuming for a lack of Javanese language use in home context and high support for various resources of English from books, movies, and games.

4. Findings and Discussions

Our findings straightforwardly address various practical ways of incorporating learner corpora in classrooms. We will focus on providing ample of examples for Javanese and English language classes.

4.1. The incorporation of learner corpora for teaching Javanese

Our corpus on the Javanese production contains 1.044 essays, 108.828 word tokens, and 10.769 word types. The first linguistic evidence we will tackle is from the word list. Figure 3 below indicates 10 most frequent words in learners' Javanese essays.



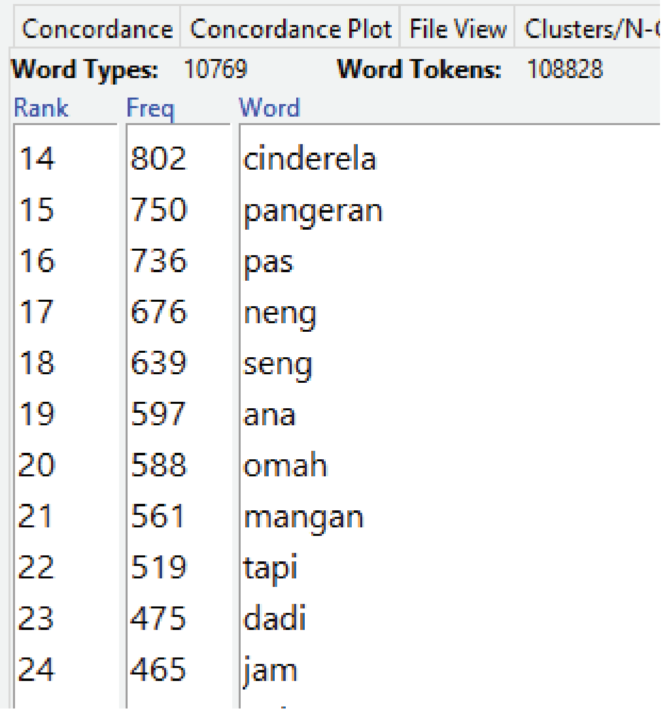
Word Types: 10769 Word Tokens: 108828

Rank	Freq	Word
1	5888	aku
2	3146	cinderella
3	2979	lan
4	1969	ing
5	1528	dolanan
6	1276	iku
7	1239	sing
8	1171	nang
9	1143	ibu
10	930	ambek

Figure 3: Word Frequency in Javanese data.

Figure 3 reveals that subjective pronoun *aku* 'I', name *cinderella*, conjunction *lan* 'and', noun *dolanan* 'toy', demonstrative *iku* 'that', preposition *nang* 'which', noun *ibu* 'mother' and preposition *ambek* 'with' are the 10 most frequent words used in the essays. Among these types, grammatical bin of conjunction, demonstrative and preposition take the most parts turning over the content words of noun and pronoun. It is surprising that Javanese verbs do not occur regularly as opposed to the writing tasks that were to tell school activities and favorite toys and retell a fairy tale. Looking at the nature of the task, we had expected to find verbs on the list. As a response to this particular finding, we may want to refer to Ruoff (1981) and Pregel and Rickheit (1987), both in Segbers and Schroeder (2016).Their study with German school children confirmed that children’s vocabularies contain about 55% nouns, about 35% verbs and 10% adjectives, based on language production by 6- to 10-year old children. It implies that the lack of Javanese verbs in our leaner corpora indicate the lack of teaching emphasis on this part of speech. This empirical data is, indeed, essential for teachers of Javanese language in either evaluating teaching materials or preparing for the incoming ones. However, analyzing the development of parts of speech distributions with a growing vocabulary size is challenging due to various factors, according to Segbers and Schroeder (2016). With the help of corpus analysis, we believe teachers, in this very specific context, will be able to overcome the challenge of Javanese vocabulary enrichment as well as to avoid monotonous teaching materials.

Engaging in this issue, we may want to relocate the first verb that occurs from the word list to ensure our understanding toward the development of Javanese vocabularies by our learners (see Figure 4).



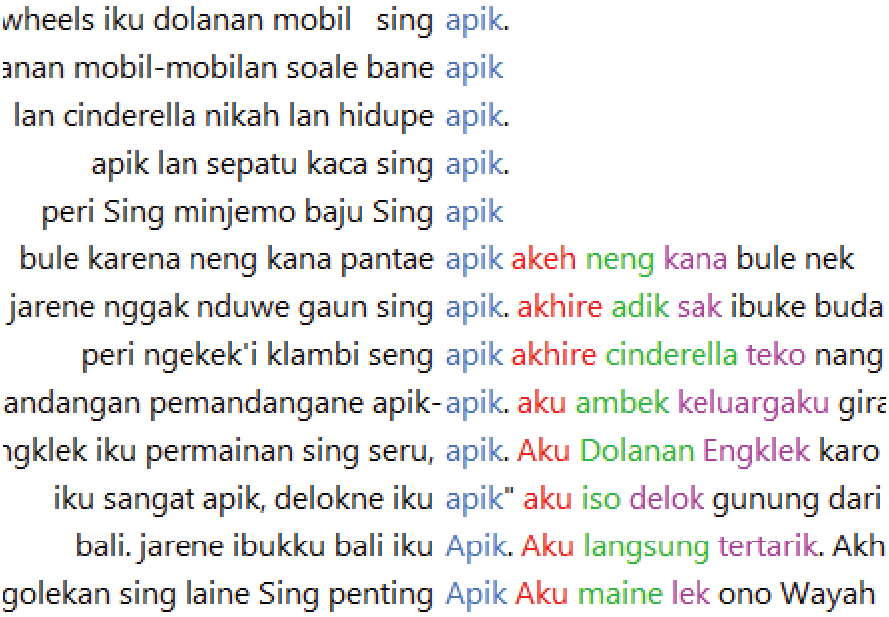
Word Types: 10769 **Word Tokens:** 108828

Rank	Freq	Word
14	802	cinderela
15	750	pangeran
16	736	pas
17	676	neng
18	639	seng
19	597	ana
20	588	omah
21	561	mangan
22	519	tapi
23	475	dadi
24	465	jam

Figure 4: Locating Javanese verb.

Figure 4 above tells us that the first verb of Javanese occurred is ‘mangan’. It comes at the 21st rank with 561 occurrences. This linguistic evidence could inform teachers, in particular, to pay more attention to expose more verbs and probably their synonyms. Again, this is one of the practical implications of how corpus evidence can contribute to teaching evaluation and preparation. In other words, by looking more thoroughly to our corpus data, we can evaluate how extensive our previous learning materials could cover especially on the range of content words and see the impact of it to the learners' Javanese language development for the purpose of preparing the next materials.

Through the lens of our corpus, it is also possible to capture code-mixing practices commonly occurred in bilingual speech production. It is the situation where a speaker mixes two languages in any level of their production as shown in Figure 5 below.



```
wheels iku dolanan mobil   sing apik.
anan mobil-mobilan soale bane apik
 lan cinderella nikah lan hidupe apik.
        apik lan sepatu kaca sing apik.
     peri Sing minjemo baju Sing apik
   bule karena neng kana pantae apik akeh neng kana bule nek
 jarene nggak nduwe gaun sing apik. akhire adik sak ibuke buda
       peri ngekek'i klambi seng apik akhire cinderella teko nang
andangan pemandangane apik- apik. aku ambek keluargaku gira
igklek iku permainan sing seru, apik. Aku Dolanan Engklek karo
    iku sangat apik, delokne iku apik" aku iso delok gunung dari
      bali. jarene ibukku bali iku Apik. Aku langsung tertarik. Akh
golekan sing laine Sing penting Apik Aku maine lek ono Wayah
```

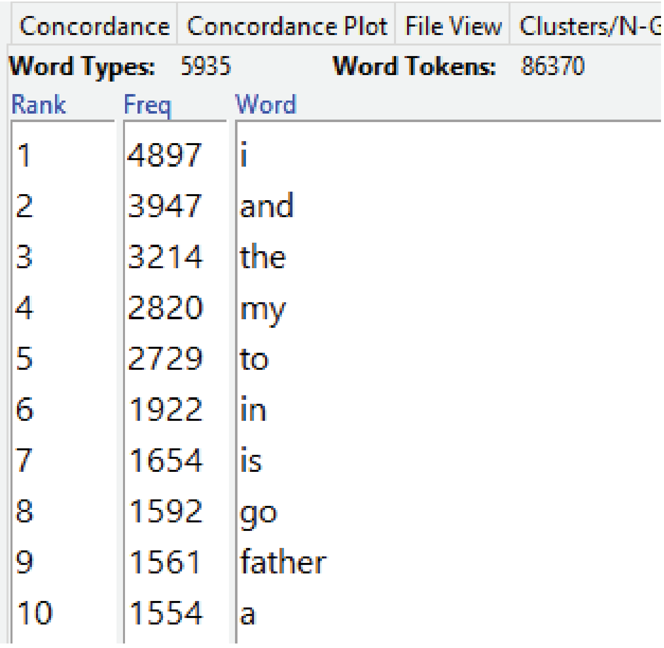
Figure 5: Concordances containing code-mixing practices.

Here, we can clearly see the Indonesian words of *mobil, mobil-mobilan, nikah, hidupe, minjemo, karena, pantae, gaun, pemandangan, permainan, sangat, maine, tertarik,* etc being mixed with Javanese words. These productions are simply unique especially on the nature of the relationship between lexical (word) and morphological development (Segbers & Schroeder, 2016). This is illustrated by the production of *minjemo, maine,* and *sangat apik*. There were seen a mix between Indonesian words with Javanese affixes. In other words, learners produce Indonesian word in Javanese morphological construction to look like 'Javanese'. It seems to us that this kind of linguistic practice has become undeniable in everyday language use of bilingual speakers. However, this corpus evidence can give a significant input for teachers, for example, to list most common Indonesian words in learners' Javanese essays, then to find their equivalence in Javanese, introduce how to use them, and emphasize them in the classroom.

Teachers of Javanese language can definitely do other kinds of observations using our learner corpora for various purposes, apart from these abovementioned preliminary analyses. Now, we will put forward some ideas of incorporating corpora for teaching English.

4.2. Using learner corpora for teaching English

Our corpus on the English production contains 980 essays, 86.370 tokens, and 5.935 word types. We will begin with evidence from word list in Figure 6 below.



Word Types: 5935 **Word Tokens:** 86370

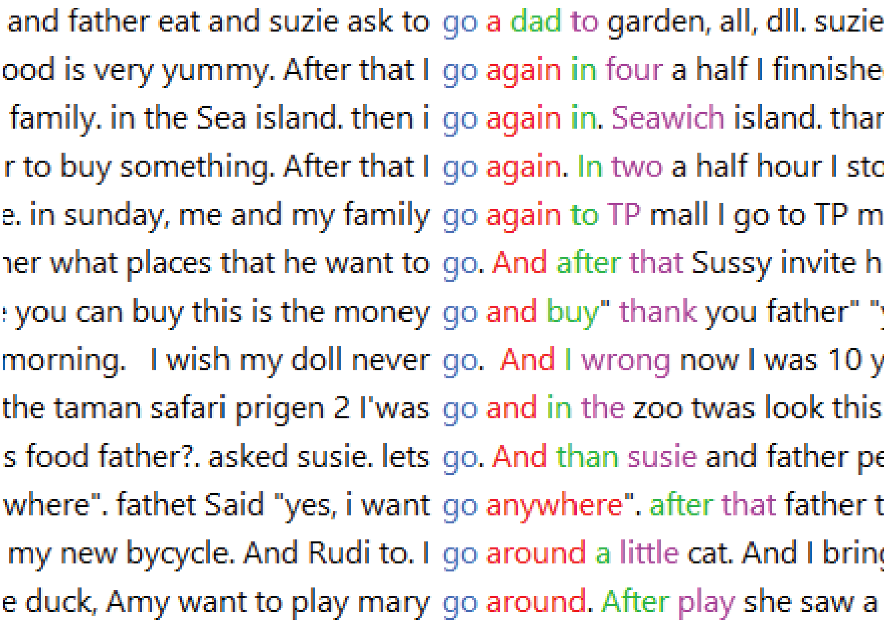
Rank	Freq	Word
1	4897	i
2	3947	and
3	3214	the
4	2820	my
5	2729	to
6	1922	in
7	1654	is
8	1592	go
9	1561	father
10	1554	a

Figure 6: Word list of English corpus.

The 10 most frequent words in English data are subjective pronoun *I*, conjunction *and*, definite article *the*, possessive *my,* preposition *to* and *in,* auxiliary verb *is*, verb *go*, noun *father*, and indefinite article *a.* Grammatical bins are still dominating. Interestingly, verb *go* occurs in the 8th rank with 1592 occurrences. Using this evidence, teachers can examine more thoroughly into how learners use the verb *go.*

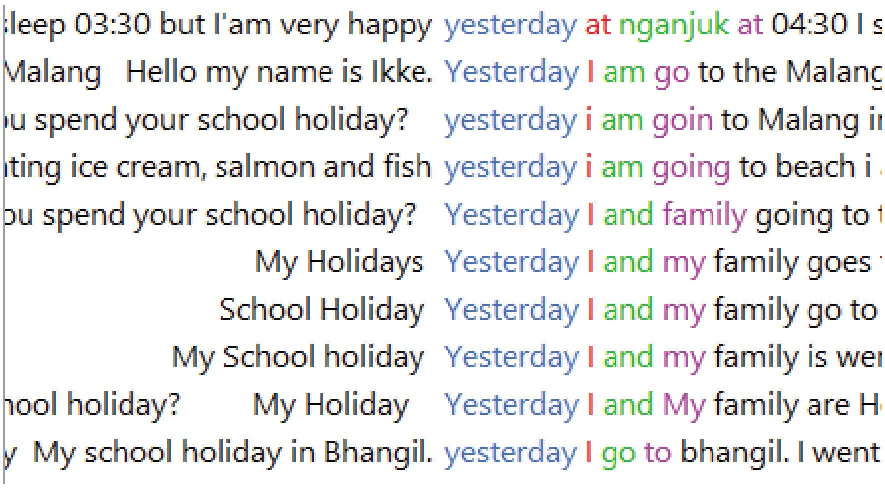
This corpus evidence has convinced us that our learners seem to be very familiar with this type of verb, from the way they use it in context. This is a starting point where teachers can continue to check the acquisition of grammatical construction of English, such as tenses, person features or so-called Phi features, and numbers as the followings.

Figuring out sentences in past forms can be done by placing a keyword of *yesterday* or any other adverbs of past time on the corpus tool. Figure 8 exemplifies concordances of past forms where we can see our learners were mostly still inaccurate in constructing past tense from the way they situated *yesterday* with the verbs that follows, such as in *Yesterday I am go, Yesterday I am going, Yesterday I and my family goes,* etc. These corpus evidences function not only as empirical data for teachers to evaluate the effectiveness of the previous materials, but more importantly as practical guidance for them to select what grammatical patterns to focus on, in the new materials. In complementing



```
and father eat and suzie ask to  go a dad to garden, all, dll. suzie
ood is very yummy. After that I  go again in four a half I finnishe
 family. in the Sea island. then i  go again in. Seawich island. thar
r to buy something. After that I  go again. In two a half hour I stc
e. in sunday, me and my family  go again to TP mall I go to TP m
ner what places that he want to  go. And after that Sussy invite h
: you can buy this is the money  go and buy" thank you father" "
morning.   I wish my doll never  go.  And I wrong now I was 10 y
the taman safari prigen 2 I'was  go and in the zoo twas look this
s food father?. asked susie. lets  go. And than susie and father pe
where". fathet Said "yes, i want  go anywhere". after that father t
 my new bycycle. And Rudi to. I  go around a little cat. And I brin
e duck, Amy want to play mary  go around. After play she saw a
```

Figure 7: Concordances of the verb *go*.

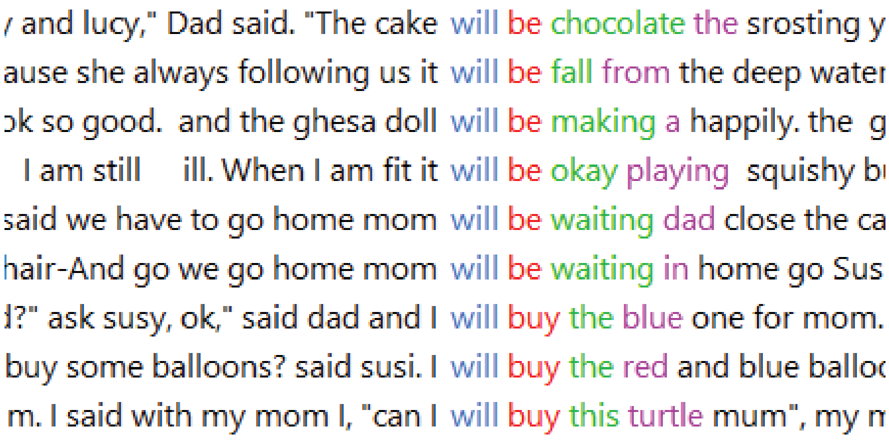


```
leep 03:30 but I'am very happy  yesterday at nganjuk at 04:30 I s
Malang   Hello my name is Ikke.  Yesterday I am go to the Malang
u spend your school holiday?     yesterday i am goin to Malang in
ting ice cream, salmon and fish  yesterday i am going to beach i
ou spend your school holiday?    Yesterday I and family going to t
                    My Holidays  Yesterday I and my family goes
                 School Holiday  Yesterday I and my family go to
              My School holiday  Yesterday I and my family is wer
ool holiday?         My Holiday  Yesterday I and My family are H
y  My school holiday in Bhangil. yesterday I go to bhangil. I went
```

Figure 8: Concordances of Past Tense.

our understanding toward the acquisition of English tense structure of our learners, we investigate the instance of future tense in Figure 9 as follows.

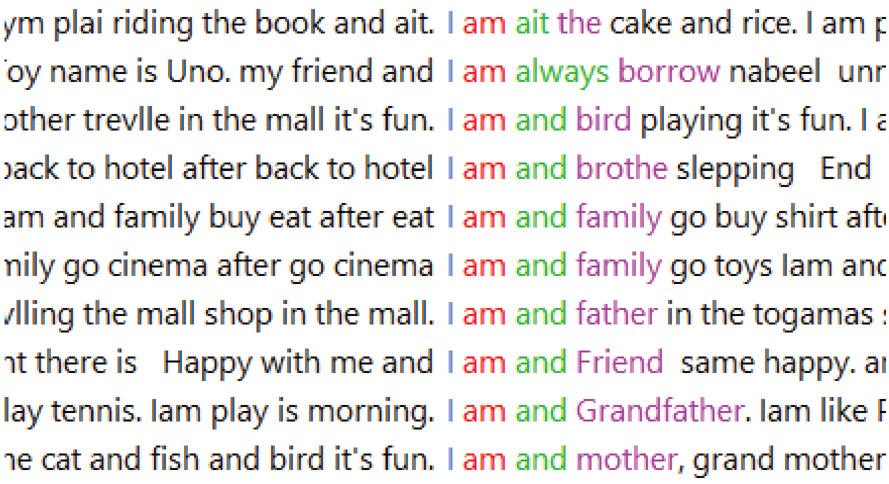
Compared to the learners' productions of past tense in Figure 8, their productions of future tense were better in a way that they accurately placed infinitive (verb 1) after modal auxiliary *will* and used *be* for nominal sentences as in *will be chocolate, will be ok*. It was also grammatically acceptable even when constructing future progressive tense, such as in *will be making, will be waiting,* even though we still find inaccurate form in *will be fall*. Teachers may want to focus on having a closer examination on error patterns that the learners produced to be highlighted in class discussion. In this circumstance, we agree to Granger (2003) in claiming that learner corpora are useful to highlight what



/ and lucy," Dad said. "The cake will be chocolate the srosting y
ause she always following us it will be fall from the deep water
ok so good. and the ghesa doll will be making a happily. the g
I am still ill. When I am fit it will be okay playing squishy bu
said we have to go home mom will be waiting dad close the ca
hair-And go we go home mom will be waiting in home go Sus
d?" ask susy, ok," said dad and I will buy the blue one for mom.
buy some balloons? said susi. I will buy the red and blue balloo
m. I said with my mom I, "can I will buy this turtle mum", my m

Figure 9: Concordances of future tense.

areas of the grammar can usefully be taught to learners. Learners can also observe their production by themselves. In this way, teachers can present the corpus data and operate the corpus tool in front of the class. Having authentic data to discuss, we believe language learning will be so much engaging.

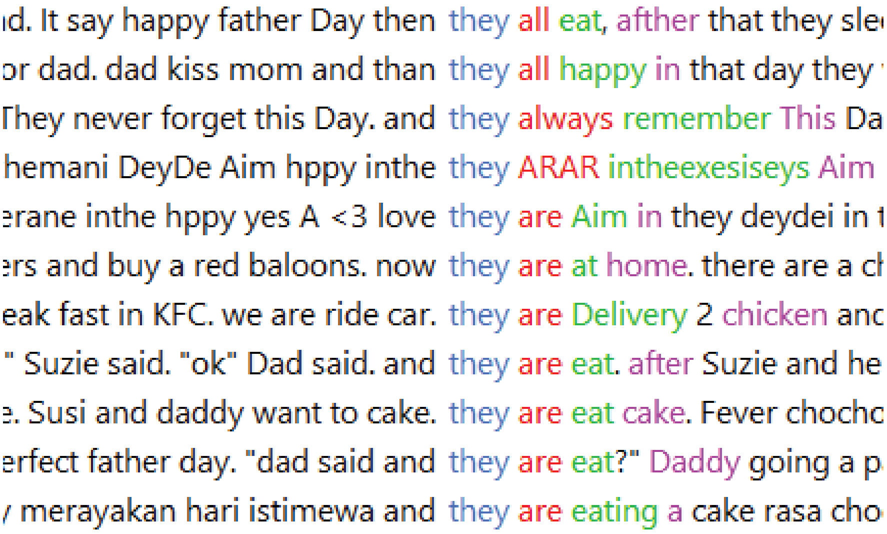


ym plai riding the book and ait. I am ait the cake and rice. I am p
oy name is Uno. my friend and I am always borrow nabeel unr
other trevlle in the mall it's fun. I am and bird playing it's fun. I a
back to hotel after back to hotel I am and brothe slepping End
am and family buy eat after eat I am and family go buy shirt aft
nily go cinema after go cinema I am and family go toys Iam anc
lling the mall shop in the mall. I am and father in the togamas s
nt there is Happy with me and I am and Friend same happy. ai
lay tennis. Iam play is morning. I am and Grandfather. Iam like F
ne cat and fish and bird it's fun. I am and mother, grand mother

Figure 10: Concordances of present progressive.

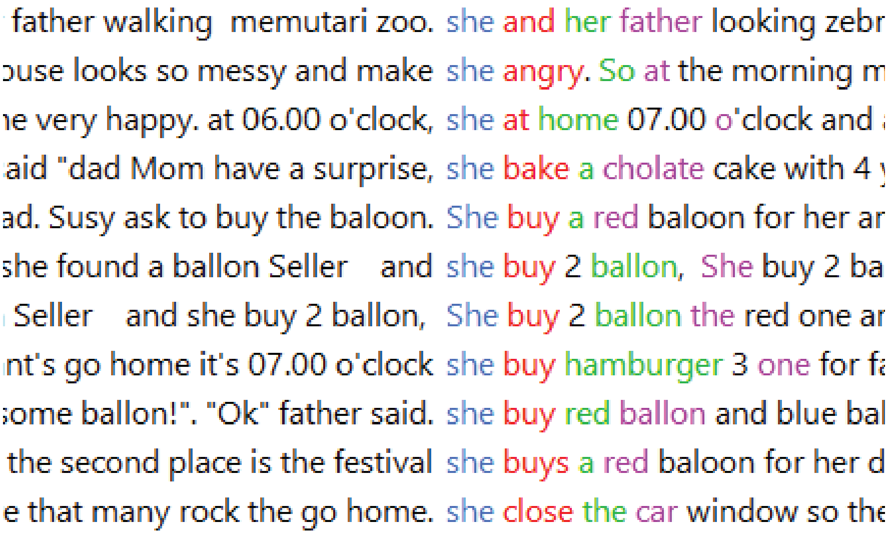
Concordances in Figure 10 enable us to see another developmental error in the production of present progressive most possibly as a result of mother tongue influence where Indonesian and/or Javanese do not posses, not only inflectional suffix –ing to mark progressivity but also subject-verb agreement. Looking more carefully at these sentences, teachers can put more creative presentations when coming into this particular tense feature. It means corpora can also help teachers improve teaching/learning strategies.

Tense features of English appear to be complex especially for young learners. It is even more complex for those whose background language, mainly L1, is structured differently, as in the case of Indonesian. Apart from tenses, person features of English exhibit different pattern from Indonesian in which subject-verb agreement applies very firmly in the former language. Through Figure 11 and 12 below we briefly compare how learners generate verb forms in relation to the behavior of third person plural and third person singular subjects.



```
id. It say happy father Day then  they all eat, afther that they sle
or dad. dad kiss mom and than  they all happy in that day they
They never forget this Day. and  they always remember This Da
hemani DeyDe Aim hppy inthe  they ARAR intheexesiseys Aim
erane inthe hppy yes A <3 love  they are Aim in they deydei in t
ers and buy a red baloons. now  they are at home. there are a cl
eak fast in KFC. we are ride car.  they are Delivery 2 chicken anc
" Suzie said. "ok" Dad said. and  they are eat. after Suzie and he
e. Susi and daddy want to cake.  they are eat cake. Fever chochc
erfect father day. "dad said and  they are eat?" Daddy going a p
/ merayakan hari istimewa and  they are eating a cake rasa cho
```

Figure 11: Concordances of *they*.

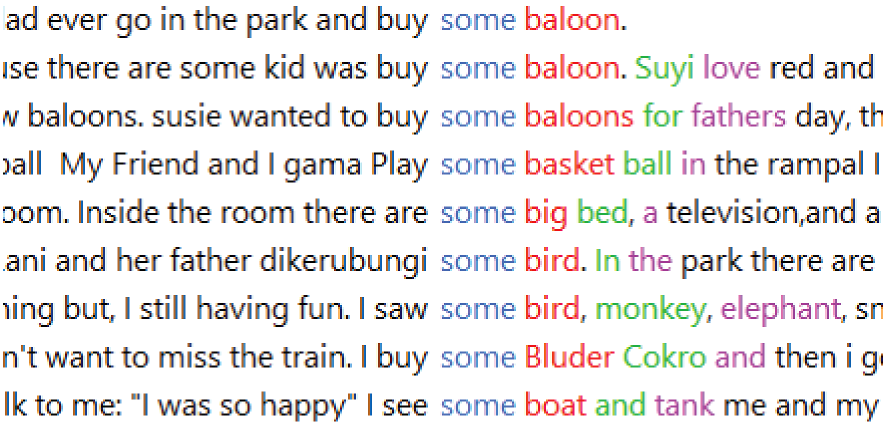


```
father walking  memutari zoo.  she and her father looking zebr
ouse looks so messy and make  she angry. So at the morning m
ne very happy. at 06.00 o'clock,  she at home 07.00 o'clock and
aid "dad Mom have a surprise,  she bake a cholate cake with 4 y
ad. Susy ask to buy the baloon.  She buy a red baloon for her ar
she found a ballon Seller    and  she buy 2 ballon,  She buy 2 ba
Seller    and she buy 2 ballon,  She buy 2 ballon the red one ar
nt's go home it's 07.00 o'clock  she buy hamburger 3 one for fa
some ballon!". "Ok" father said.  she buy red ballon and blue bal
the second place is the festival  she buys a red baloon for her d
e that many rock the go home.  she close the car window so the
```

Figure 12: Concordances of *she*.

Subject *they* in Figure 11 was mostly followed by the correct auxiliary verb *are* as well as to the verb form of *remember* and *eat.* In contrast, the subject *she* most frequently disagree to its verbs as it has to be attached by inflectional suffix –*s*. Again, we believe this finding is valuable for teachers in a sense that they can present this corpus data directly to their learners as to show which forms are accurate and which are not.

Number features deem to be basic in English grammatical constructions. However, it remains challenging for Indonesian learners as in this language number does not require any modification to a noun it modifies. To look at how singular and plural noun behave in our corpus data more precisely, we present the corpus evidence by firstly inputting the word *some* and *many* as keywords. The results are in Figure 13 and 14 below.



ad ever go in the park and buy some baloon.

ıse there are some kid was buy some baloon. Suyi love red and

v baloons. susie wanted to buy some baloons for fathers day, th

ɔall My Friend and I gama Play some basket ball in the rampal I

ɔom. Inside the room there are some big bed, a television,and a

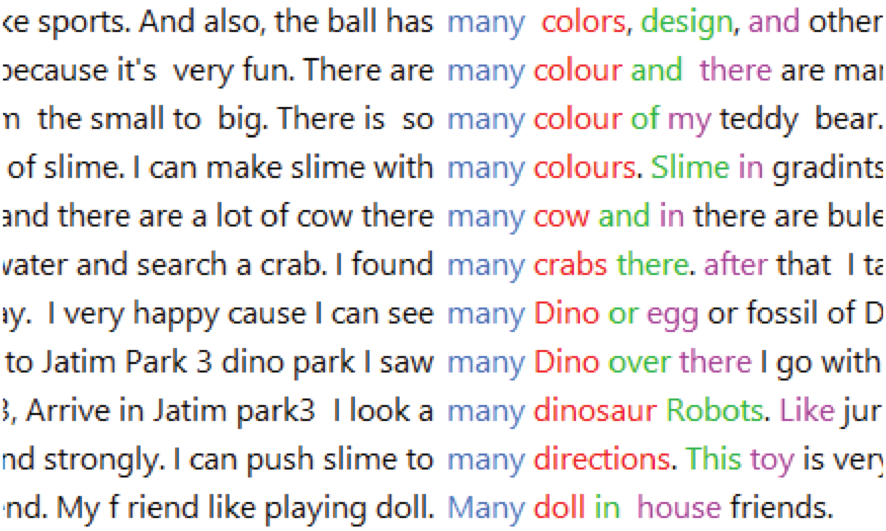
ani and her father dikerubungi some bird. In the park there are

ıing but, I still having fun. I saw some bird, monkey, elephant, sn

n't want to miss the train. I buy some Bluder Cokro and then i g

lk to me: "I was so happy" I see some boat and tank me and my

Figure 13: Concordances of *some* expressing plurality.



ce sports. And also, the ball has many colors, design, and other

ɔecause it's very fun. There are many colour and there are mai

n the small to big. There is so many colour of my teddy bear.

of slime. I can make slime with many colours. Slime in gradints

ınd there are a lot of cow there many cow and in there are bule

vater and search a crab. I found many crabs there. after that I ta

ıy. I very happy cause I can see many Dino or egg or fossil of D

to Jatim Park 3 dino park I saw many Dino over there I go with

3, Arrive in Jatim park3 I look a many dinosaur Robots. Like jur

nd strongly. I can push slime to many directions. This toy is ver

nd. My f riend like playing doll. Many doll in house friends.

Figure 14: Concordances of *many*.

We all can see that some learners produced accurate nouns in regard to its modifier, such as *some balloons, many colors,* and *many crabs*. Yet, some other productions were still incorrect, such as *some bird, some balloon,* and *many doll.* As previously stated, such data can be invaluable input for teachers to do a brief statistical analysis toward the number of accurate production as to see what teaching/learning activities work best for learners.

Along these lines, as learner corpora include database information of particular learners, they usually contain first language background, age, gender, proficiency level and length of time learning the target language which, in turn, become useful when interpreting developmental patterns of the learners (Mccarthy, 2016).

5. Conclusions and Suggestions

To conclude, here are several attempts on how our learner corpora can practically facilitate teachers in language classrooms; (1) teachers can have authentic evidence of learners' language production that may reflect the quality of teaching and learning, (2) these corpus evidences can help teachers see the extent of learners' vocabulary acquisition and development as well as to guide teachers to move out from monotonous vocabulary teaching into more creative and engaging ones, (3) these corpus data can show teachers the most frequent grammatical errors learners had produced including spelling, tense, word structures etc. as for teachers to select what features should come first and what to highlight in the class, (4) these learner corpora can also be employed as databases to monitor learners' language acquisition and development which is essential for curriculum design, and (5) these corpora can be utilized as a starting point for educational policy makers especially when designing learning outcomes (e.g. the vocabulary size should be acquired or mastered by learners of each grade).

As knowing the effectiveness of corpus use in teaching and learning, we highly value the importance of conducting workshops on how to engage corpus into classroom where teachers will be trained to build learner corpora on their own, to explore, and to exploit them for various teaching/learning purposes.

Acknowledgement

We are thankful for Rebecca Lurie Starr, the main author's PhD Supervisor, for her constant guidance during data collection. The appreciation also goes to LPDP (Lembaga Pengelola Dana Pendidikan) for a research fund awarded to the main author.

References

[1] Apriana, A., Kadarisman, E., & Yaniafari, R.P. (2017). *Building a corpus on children's written production*. (Unpublished research report, State University of Malang, Indonesia).

[2] Botley, S. P. (2014). Argument structure in learner writing: A corpus-based analysis using argument mapping. *Kajian Malaysia*, *32*(1), 45–77.

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

[4] Granath, S. (2018). Who benefits from learning how to use corpora? In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 47–65). The Netherlands: John Benjamin Publishing Company.

[5] 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.

[6] Hana, J., Rosen, A., Stindlova, B., & Stepanek, J. (2014). Building a learner corpus. *Language Resources & Evaluation*, *48*, 741–752. http://doi.org/10.1007/s10579-014-9278-z

[7] Johns, T. F. (2002). Data-driven learning: The perpetual challenge. In *Teaching and Learning by Doing Corpus Analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19–24 July, 2000,* B. Kettemann & G. Marko (eds), 107–117. Amsterdam: Rodopi.

[8] Kennedy, G. (1998). An introduction to corpus linguistics. London: Longman.

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

[10] Mccarthy, M. (2016). Putting the CEFR to good use: Designing grammars based on learner-corpus evidence. *Language Teaching2*, *49*(1), 99–115. http://doi.org/10.1017/S0261444813000189

[11] McEnery, T. and Wilson, A. (1996) *Corpus linguistics*. Edinburgh: Edinburgh University Press.

[12] Meunier, F., & Gouverneur, C. (2018). New types of corpora for new educational challenges. In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 179–201). The Netherlands: John Benjamin Publishing Company.

[13] Meunier, F., & Littre, D. (2013). Tracking learners' progress: Adopting a dual "Corpus cum Experimental Data" Approach. *The Modern Language Journal*, *97*, 61–76. http://doi.org/10.1111/j.1540-4781.2012.01424.x

[14] Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In: Sinclair, J. (ed.), *How to use corpora in language teaching* (pp. 125–152). Amsterdam: John Benjamins.

[15] O'Keeffe, A., McCarthy, M. and Carter, R. (2007) *From corpus to classroom*. Cambridge: Cambridge University Press.

[16] 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

[17] 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

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

[19] Römer, U. (2018). Corpus research and practice: What help do teachers need and what can we offer? In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 83–98). The Netherlands: John Benjamin Publishing Company.

[20] 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.

[21] Shirato, J., & Stapleton, P. (2007). Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan. *Language Teaching Research*, *11*(4), 393–412.

[22] Sinclair, J. (1991) *Corpus, Concordance and Collocation.* Oxford: Oxford University Press.

[23] Sundh, S. (2016). A corpus of young learners' English in the Baltic region - Texts for studies on sustainable development. *Discourse and Communication for Sustainable Education*, *7*(2), 92–105. http://doi.org/10.1515/dcse-2016-0018.

[24] Timmis, I. (2015). *Corpus Linguistics for ELT: Research and Practice*. London and New York: Routledge.