International Dialogues on Education, 2019, Volume 6, Number 2, pp. 83-98
ISSN 2198-5944

*Ira Rasikawati (Indonesia)*

# Corpus-Based Data-Driven Learning to Augment L2 Students' Vocabulary Repertoire

***Abstract:*** *Corpus-based data-driven learning (DDL) is an inductive instructional approach using computer-generated concordances. It provides students with the opportunity to analyze different language forms across contexts found in the concordance output. The idea of engaging students to discover the language rules and patterns from authentic learning materials is central to the theory of inquiry-based learning. Despite the robust research support, however, DDL has not been widely adopted, in part because of a dearth of practical and specific recommendations for teachers. More studies are needed to corroborate the claim that the approach can promote the development of different language learning areas effectively. This article synthesizes relevant theories and research findings on the use of DDL for second language instruction and illuminates the understanding of how corpus-based vocabulary instructional strategies may work in English for Academic Purposes (EAP) courses in non-English speaking countries. The study recommendations include a corpus-based DDL framework to expand students' vocabulary and suggestions for future research*
***Keywords:*** *EAP, second language, vocabulary, corpora, DDL, inquiry-based learning*

* * *

***摘要*** *(Ira Rasikawati: 基于语料库的数据驱动学习，以扩大二年级学生的词汇量）：基于语料库的数据驱动学习（DDL）是一种使用计算机生成的一致性的归纳式教学法。它为学生提供了在一致性输出中所发现的跨越上下文分析不同语言形式的机会。让学生从真实的学习材料中发现语言的规则和模式的这一想法是探究式学习理论的中心。尽管具有强大的研究支持，但DDL尚未被广泛运用，部分原因是由于缺少针对教师的实用性的及专业的建议。更多的研究需要来证实这种方法可以有效地促进不同语言学习领域的发展。本文综合了有关在第二语言教学中使用DDL的相关理论和研究成果，并阐明了在基于语料库的词汇教学策略下，在非英语国家的以学术英语（EAP）为目的的工作中如何发挥作用的理解。研究建议包含了一个基于语料库的DDL框架，以扩展学生的词汇量和对未来研究的提议。*
***关键词****：EAP，第二语言，词汇，语料库，DDL，探究式学习*
***摘要*** *(Ira Rasikawati: 基於語料庫的數據驅動學習，以擴大二年級學生的詞彙量）：基於語料庫的數據驅動學習（DDL）是一種使用計算機生成的一致性的歸納式教學法。它為學生提供了在一致性輸出中所發現的跨越上下文分析不同語言形式的機會。讓學生從真實的學習材料中發現語言的規則和模式的這一想法是探究式學習理論的中心。儘管具有強大的研究支持，但DDL尚未被廣泛運用，部分原因是由於缺少針對教師的實用性的及專業的建議。更多的研究需要來證實這種方法可以有效地促進不同語言學習領域的發展。本文綜合了有關在第二語言教學中使用DDL的相關理論和研究成果，並闡明了在基於語料庫的詞彙教學策略下，在非英語國家的以學術英語（EAP）為目的的工作中如何發揮作用的理解。研究建議包含了一個基於語料庫的DDL框架，以擴展學生的詞彙量和對未來研究的提議。*
***關鍵詞****：EAP，第二語言，詞彙，語料庫，DDL，探究式學習*

* * *

***Zusammenfassung*** *(Ira Rasikawati: Korpusbasiertes datengesteuertes Lernen zur Erweiterung des Wortschatzrepertoires der L2-Studierenden): Corpus-based Data Driven Learning (DDL) ist ein induktiver Lehransatz unter Verwendung computergenerierter Konkordanzen. Es bietet den Studenten die Möglichkeit, verschiedene Sprachformen kontextübergreifend zu analysieren, die in der Konkordanz-Ausgabe zu finden sind. Die Idee, die Schüler dazu anzuhalten, die Sprachregeln und -muster aus authentischen Lernmaterialien zu entdecken, ist von zentraler Bedeutung für die Theorie des forschungsbasierten Lernens. Trotz der robusten*

*Forschungsunterstützung ist DDL jedoch nicht weit verbreitet, auch weil es an praktischen und spezifischen Empfehlungen für LehrerInnen mangelt. Weitere Studien sind erforderlich, um die Behauptung zu untermauern, dass der Ansatz die Entwicklung verschiedener Sprachlernbereiche wirksam fördern kann. Dieser Artikel fasst relevante Theorien und Forschungsergebnisse über den Einsatz von DDL im Zweitsprachenunterricht zusammen und beleuchtet das Verständnis dafür, wie korpusbasierte Vokabularlehrstrategien in Englisch für akademische Zwecke (EAP) Kurse in nicht-englischsprachigen Ländern funktionieren können. Die Studienempfehlungen beinhalten einen korpusbasierten DDL-Rahmen zur Erweiterung des Wortschatzes der Studierenden und Vorschläge für zukünftige Forschungen.*
***Schlüsselwörter:*** *EAP, Zweitsprache, Vokabular, Korpora, DDL, forschungsbasiertes Lernen*

* * *

***Аннотация*** *(Ира Разикавати: Обучение на основе данных для расширения словарного запаса студентов педагогических направлений подготовки): Обучение на основе данных – индуктивный подход с применением компьютерных технологий и индексов компьютерной лингвистики. Данный метод дает возможность студентам анализировать различные языковые формы в любых контекстных ситуациях Эти языковые формы заданы в системе индексов. Идея научить студентов пользоваться данным методом для освоения определенных языковых правил и образцов на примере работы с аутентичным материалом имеет центральное значение для теории обучения на основе данных. Тем не менее стоит отметить, что несмотря на мощную «раскрутку» данного метода, он еще не получил широкого применения, поскольку на данный момент ощущается недостаток в практических и других специальных рекомендациях для учителей по внедрению метода. Необходимо проводить дальнейшие исследования для того, чтобы подтвердить высказанное положение о том, что данный подход стимулирует освоение различных участков языкового пространства. В данной статье рассматриваются наиболее значимые теории и практические результаты использования метода обучения на основе данных при освоении второго языка; в работе также показано, как стратегии обучения лексике английского языка на основе корпусной лингвистики (английский язык для академических целей) могут применяться в тех странах, где английский язык не является государственным. Разработанные рекомендации опираются на данный метод и направлены на решение задачи по расширению словарного запаса студентов; в статье указываются также перспективы дальнейшего изучения данного вопроса.*
***Ключевые слова:*** *английский язык для академических целей, второй язык, вокабуляр, корпус, обучение на основе данных, обучение через исследование*



# Background

Acquiring a new language is in many respects challenging for language learners. Vocabulary acquisition can be especially demanding for learners of English as a Foreign Language (EFL), whose learning contexts are non-English speaking countries. Unlike second language (L2) learners who receive rich language exposure in school, EFL learners lack sufficient input in their learning environment (Kojic-Sabo & Lightbown, 1999). Moreover, EFL instructional texts and teacher education has paid little attention to vocabulary acquisition (Hunt & Beglar, 2005). The difficulty of acquiring a large vocabulary size becomes more apparent as students enter higher education. The increasing status of English as the dominant language of science has put greater demands for students to read English references for their study purposes. EFL instruction aiming at improving students' ability to read academic texts becomes paramount in various higher education settings.

Learning English for study purposes can be daunting for L2 including EFL students as it requires students' ability to think about academic content and convey abstract ideas in the language they are still learning (Nagy & Townsend, 2012; Snow & Uccelli, 2009). Students need to develop not only the breadth but also depth of their vocabulary knowledge. An immense amount of vocabulary is fundamen-

tal for them to read English texts independently. According to Nation (2006), students need to know 98% of the running words in a text for adequate comprehension, and even 98% of word coverage may not suffice for easy understanding of a non-fiction text (Carver, 1994; Kurnia, 2003). Nation (2006) estimates that students need to have a vocabulary size of 8,000 to 9,000 word-family for written text comprehension and of 6,000 to 7,000 for spoken text comprehension.

Mature readers additionally need to increase the depth of their vocabulary knowledge, which necessitates the mastery of a wide range of vocabulary knowledge: polysemous meanings; collocations; word uses and forms (Schmitt, 2014). Knowing vocabulary, therefore, entails having a large vocabulary size and understanding of the nuances of words' meaning for both receptive and productive purposes. Considering the complexity of learning vocabulary knowledge, developing student vocabulary for academic purposes requires word learning strategies beyond the generic approaches. Researchers and practitioners have examined different ways to find new possibilities for facilitating L2 learning more successfully in the classroom by incorporating digital technology.

Information and communication technology (ICT) is a common term to refer to the use of digital technology in language teaching and learning (Evans, 2009). ICT has incentivized EFL students in that it provides them with a tool to make learning more productive and self-regulated. Effective integration of ICT in the classrooms has been purported to bring about development in various learning skills such as "reasoning and problem solving, learning how to learn and creativity"; broaden and deepen learning; and increase interest in learning activities (Eadie, 2001, p. 28). The use of technology for language learning has evolved from what is termed computer-based training (CBT) and computer-assisted language learning (CALL) (Farr & Murray, 2016) to the highly innovative corpus-based approach following the growth in corpus linguistics research (Warren, 2016).



Research in L2 teaching and learning has been appealed to by the data-driven learning (DDL) approach, a term coined by Tim John (1991) to refer to the application of corpora to investigate language use. This approach, rooted in the principles of discovery or inquiry learning, allows the practice of learning English through the discovery of language patterns from the data presented in the corpus – an extensive collection of authentic texts either spoken or written on a computer for language analysis. A growing body of evidence from empirical studies in the second language (L2) classrooms indicates that corpora offer abundant opportunities for data-driven learning (DDL) potent for students' language development. Although perceived as a promising L2 learning approach, DDL appeared to encounter stagnancy on the classroom path. Notwithstanding the availability of free online corpora and classroom guides, DDL has not been "part of mainstream teaching practice" (Boulton, 2010, p. 534).

The primary purpose of this article is to present a literature review on the theoretical constructs and empirical findings which rationalize the practice of corpus-based vocabulary instruction in a reading English for Academic Purposes (EAP) classroom. Prior to the analysis of previous studies reporting the use of corpus-based DDL in the second language classrooms, the article reviews the discovery learning or inquiry-based learning theory which explains the philosophical constructs of the DDL approach. The findings of a case study in an EAP classroom in Indonesia illuminates how the perceived limitations and challenges of DDL may be addressed and informs the proposed framework for corpus-based vocabulary development. The recommendations include the practical implications for EFL classrooms and future research directions.

## Inquiry-Based Learning Revisited

The term inquiry learning is often used interchangeably with discovery learning to refer to the method of instruction which conceptual roots can be traced back to the works of the constructivists such as John Dewey and Jerome Bruner. Herbert Spencer's (1820-1903) notion of teaching students "how to think" instead of "what to think" seems to have significantly influenced Dewey (1910) in his publication of *How We Think* and Bruner (1961) in his notions of "learning how to learn" and "inquiry discovery" (Ornstein & Hunkins, 2016). Dewey (1910) asserted that students' experience and prior knowledge are valuable resources which students draw upon to make sense of new information, identify connections between past and present learning experiences, and construct solutions to solve problems. Bruner (1961) similarly believed that learning brings about powerful effects when students can discover new facts and relationships of those facts for themselves and which experiences are build up from their prior learning. Dewey's and Bruner's ideas of how knowledge is constructed become the foundation of discovery learning in that the approach engages students in pursuing their interests and questioning of the existing beliefs and assumptions.

Kirschner, Sweller, & Clark (2006) consider the pedagogy in this discovery learning umbrella as the minimally guided method in contrast to the direct instructional method which provides students with both concepts and learning strategies explicitly. Other corresponding pedagogical approaches include problem-based learning, experiential learning, and constructivist learning (Kirschner et al., 2006). Today, discovery learning appears to exist on a continuum from pure inquiry to guided inquiry. In this article, the term inquiry-based learning (IBL) refers to the learning approach that is inquiry-oriented and holds students accountable for their learning (Blessinger & Carfora, 2014). The approach the same time recognizes the teachers' role to scaffold the process of inquiry to achieve curricular goals (Coffman, 2017). Although the implementation IBL strategies can be context-bound depending on the students' needs and their learning goals, teachers play an essential role in designing structured activities that promote higher-level thinking, also known as, high-order thinking.



Hattie (2012) discusses the controversy over the level of directness necessary in instruction, suggesting that the explicit approach better promotes learning and teachers are responsible for enhancing student learning actively. He highlights the study finding of Alfieri, Brooks, Aldrich, and Tenenbaum (2011) examining the efficacy of discovery-based learning which concludes that students benefit more from instructional approaches that allow scaffolded tasks and feedback for them to construct new understanding built on their existing knowledge. In response to the critics of IBL including those of Kirschner et al. (2006), Blessinger & Carfora (2014) suggest that a significant number of studies has persuasively argued for IBL as a promising instructional strategy when properly designed and implemented. IBL can promote student engagement, motivation, autonomy, and problem-solving skills when the teacher plan the course and facilitate the learning effectively (Blessinger & Carfora, 2014, p. 5). The authors recommend that the teacher offer active and caring support to ensure a conducive learning environment and sufficient guidance for the students.

The contemporary IBL continues to be an approach to learning that is self-directed, question-driven, and problems relevant (Levy, Lameras, McKinney, & Ford, 2011). In achieving the expected learning outcomes, instructional goals, content, and practices must correspond to the learning assessment (Blessinger & Carfora, 2014). Although the students are held accountable for their learning, teachers as the subject matter experts need to provide ample support for them to engage in authentic and meaningful activities (Blessinger & Carfora, 2014; Coffman, 2017). As teachers engage students through the scaffolded learning process, the integration of digital technologies potentially enhances their learning.

Coffman (2017) asserts that the interaction between digital technologies and information improves inquiry-oriented learning. The critical elements of IBL as stated by Blessinger and Carfora (2014) – exploration and investigation; authentic inquiries using contextual and situated learning; and research-based approach – are in line with the elements of data-driven learning.

## Data-Driven Learning

The authenticity of the learning materials and students' engagement to discover the language rules and patterns from the concordance lines –list words within the contexts of each word's occurrences – are DDL attributes parallel to those of the IBL. The use of corpus linguistics – the compilation and analysis of corpora – was initially advocated by John Sinclair (Johns, 1994; Moon, 2007). The term data-driven learning (DDL) was later popularized to refer to the language learning strategy that allows students to be "language detectives" or "researchers" to explore language data (Johns, 1991). When Johns initially used DDL with his postgraduate students to improve their writing, he worked with the limited availability of concordancer compared to the present day (Boulton, 2012).

John (1991) claimed that this inductive approach benefits students more than the traditional grammar-based approach in that it allows them the opportunity to examine the linguistic patterns and generalize the rules from the language examples. The language teacher's role is to guide students with discovery strategies so they can "learn how to learn" (1991, p. 1). John's argument is relevant to the notion that L2 acquisition process can be more analytic than L1 since L2 learners have acquired a language system and received instruction in morphosyntactic rules which enable them to analyze a large unit of meanings into smaller segments (Wang, 2016). Sinclair (2004) suggests that for students to derive the most benefits from corpus resources, they need facilitation from teachers to investigate the complex corpus data. Therefore, productive corpus-based discovery learning activities as Keck (2004) put it, requires the collaborative efforts of teachers and students to select appropriate corpora, perform productive queries and interpretation of data to make generalizations of the language use.



Arguing against the critics that DDL might not be appropriate for a broader audience of language learners, Shaw (2011) states that the latest development has made corpus-based DDL possible to cater to different types of learners. DDL is presently more popular among language teaching practitioners as the development of computing resources has made corpora available online for language learners such as the free access to the Corpus of Contemporary American English (COCA) (Davies, 2008–), the British National Corpus (BNC) (BNC, 2018), and BYU-BNC (Davies, 2004–). Corpus linguistics has largely influenced language instruction today in different ways, including the development of corpus-informed materials, corpus-cited texts, and corpus-designed activities (Bennet, 2010). That said classroom materials have integrated language patterns and word frequency information derived from corpora; learning tasks have included analyses of authentic language samples retrieved from the corpora; and language samples in dictionaries contain more authentic corpus data.

To illustrate the implication of corpus linguistics for learning in the classroom, Johns uses the term "to cut out the middleman … to give the learner direct access to the data" (1994, p. 30). Giving the students direct access to the language data emphasizes the students' active role for their learning, the discovery nature of the method, and the authenticity of the task. O'Sullivan (2007) offers a comprehensive account of the DDL benefits from the perspective of the process-oriented approach to language learning which conclusion suggests that corpus consultation increases learners' mental activity, cognitive abilities, metacognitive knowledge, and independent learning. Reviewing what empirical studies have found

on the use of the corpus-based DDL – its efficacy and challenges – is the next important step before advancing to the appropriate framework for vocabulary development in the EFL context.

# Corpora in Language Teaching and Learning

A large body of research in L2 learning has examined the direct applications of corpus linguistics in language pedagogy. The efforts stemmed from the projects of Collins COBUILD English Language Dictionary 1987 which language data was based on the analyses of electronic corpora (Biber & Reppen, 2015; McEnery & Xiao, 2010; Szudarski, 2018). Most studies within applied linguistics after this period, especially in the subfields of English for Specific Purposes (ESP) and English for Academic Purposes (EAP) have benefited from the corpus research and analysis data (Biber & Reppen, 2015). Keck (2004) states that the use of corpora in language teaching includes the domains of corpus-based language descriptions, corpus-based language analysis in the classroom, and learner corpus analysis. This section reviews research findings assessing the corpus-based language analysis in L2 classrooms. A meta-analysis of Boulton & Cobb (2017) systematically examined previous studies on the use of corpus linguistics for L2 learning. Other empirical study findings provide evidence for the rationale and strategies for using corpus consultation to increase students' knowledge in different language areas.

***Meta-analysis results***. The most recent, possibly the first meta-analysis on the use of corpus in language learning is that of Boulton & Cobb (2017). The study offers compelling evidence supporting the use of corpus linguistics for L2 development programs. The result of 64 meta-analyzable studies from a pool of 205 studies showed large overall effects for control/experimental group comparisons ($d$ = 0.95) and for pre/posttest designs ($d$ = 1.50) (Boulton & Cobb, 2017). Although it is rather early to claim that the approach may yield as strong results in real classrooms which conditions are more complex than those in the experimental studies, the high $d$ values observed in various studies involving over 3,000 participants is quite encouraging.



In answering their research questions, Boulton & Cobb made three main conclusions. First, "DDL research is a flourishing field" with at least 205 publications reporting quantitative study findings since 2014 (2017, p. 381). Second, both the effectiveness and efficiency studies on the use of DDL to increase learners' L2 skills and knowledge yielded large effect sizes. Third, DDL showed consistent significant effects in situations where (1) the presence of native English instructors was limited; (2) courses targeted undergraduate and graduate learners as well as those of intermediate and advanced English levels; (3) computer- and paper-based concordancing were used; and (4) corpora were used either for a reference resource or learning vocabulary and lexicogrammar (2017, p. 383-385). A closer look into the individual study can provide a more explicit account of which language development area DDL may work most effectively.

***Authentic language input***. EFL learners are often disadvantaged by the lack of exposure to authentic language input compared to L1 or L2 learners. The non-native English teachers, grammar-based instructional materials, and direct instructional methods in a way constrain learners' access to the naturally occurring English such as the appropriate use of collocations and colligations. Students may be able to produce grammatically correct sentences, but their patterns are not common in English. The most significant benefit of Corpus-based DDL lies in the authenticity of the language to be analyzed by students (Clifton & Phillips, 2006; Romer, 2008). Using corpora in language classroom provides learners with nuanced language samples which usage of vocabulary, grammar, and functions are similar to those in natural settings (McEnery & Xiao, 2010).

Studies reporting the effectiveness of corpus-based DDL in undergraduate classrooms include those of Daskalovska (2015), Huang (2014), and Gordani (2012). Daskalovska (2015) compared the effectiveness of corpus-based activities and traditional activities for learning collocations among 46 first-year students in a university in the Republic of Macedonia. The study found that students in the experimental group, who used the online concordancer learned more collocations and showed better results on the test that measured their knowledge of verb-adverb collocations.

Previous empirical studies have yielded some valuable insights into the use of corpus-based DDL at the university level suggesting that concordance lines and other corpus query results provide learners with authentic samples of texts and vocabulary needed for language analysis and reference. Exposing students with how words are used in real texts rather than simplified texts tailored for L2 learning purpose potentially result in a more significant improvement in their vocabulary knowledge as well as the ability to comprehend and produce authentic language in the target language.

A plethora of other literature and research to date have supported the efficacy of corpus-based DDL to promote specific areas of language development. The approach facilitates the acquisition of English vocabulary (Karras, 2016), lexico-grammatical patterns (Huang, 2014; Liu & Jiang, 2009), and speaking fluency (Geluso & Yamaguchi, 2014). Students additionally improve their ability to use familiar words in new ways (Frankenberg-Garcia, 2012). Being engaged in corpus-based queries and analyses helps students develop their knowledge of linking adverbials in English (Boulton, 2009) and English verb-noun collocations (Chan & Liou, 2005) as well as strengthen their ability to use the passive voice (Smart, 2014).

***Attention in L2 learning.*** Other benefits claimed concern with the development of students' metacognitive and cognitive skills through inductive and deductive reasoning activities (Boulton, 2009). Students enhance their language noticing and autonomy by engaging in inquiry-based activities (Boulton, 2017; Chambers, 2007; Godwin-Jones, 2017; Yoon & Hirvela, 2004). These findings are in agreement with the notions of Schmidt (1995, 2001) that attention plays a significant role in retention and all types of learning, including in second language learning. The implications for L2 learning include learners need to pay attention to the language input and compare their utterances and those of the target language speakers. By finding clues derived from language samples, L2 learners are expected to notice how language samples occur in specific contexts and generate principles of how the target language works (Schmidt, 1995).



Considering the favorable study findings in different contexts, corpus-based DDL seems to be an approach that embraces a broader range of opportunities for students to develop both their cognitive and metacognitive skills to acquire academic language. Although most of the research results are in support of corpus consultation strategies in second language learning, the findings are also subject to caveats. The limitations and challenges identified in the studies are to be considered along with the recommendations and implications for instruction.

## Addressing the Challenges and Limitations of DDL

In addition to its benefits, Liu and Lei (2017) state a wide range of challenges and limitations of DDL based on the findings of various studies. The challenges concern with the difficulty of corpus query analysis (Boulton, 2009; Liu & Jiang, 2009); the need for intensive training for corpus analysis (Boulton, 2009; Karras, 2016; Liu & Jiang, 2009; O'Keeffe & Farr, 2003); and the paucity and/or difficulty of access to corpus search engines (Kennedy & Miceli, 2001; Kosem, 2008; Liu & Jiang, 2009). The consequence that follows, students may be reluctant to conduct their corpus-based search due to the difficul-

ty of running the query in addition to analyzing and interpreting the query results. The DDL approach also seems to favor more the higher-level learners than the lower ones (Liu & Lei, 2017; Boulton, 2009) as sufficient analytical and linguistic skills are necessary to cope with the complexity of the authentic data presented in the corpus query results (Boulton, 2009).

Although corpus-based DDL may be most appropriate for university students having intermediate to advanced English proficiency, Boulton & Cobb (2017) consider this claim arguable since limited studies have examined the use of DDL at the secondary level. Karras (2016), one of a few studies investigating the use of corpus-based DDL among secondary students, found that allowing sufficient training time led to better results in the vocabulary acquisition programs targeting students in this education level. While generally computer-based corpora may be preferred due to the ease of access and the extensive data for analysis, fears and lack of technology in the classroom are other voiced concerns that can hinder the implementation of corpus-based DDL. Recent study findings have offered salient recommendations to address the perceived challenges of using corpus-based DDL effectively.

***Paper-based concordance***. Although computer-based corpora promise a more controlled "massive contextual exposure" to the target language compared to a regular reading or listening program (Boulton, 2017, p. 483), the paper-based concordancing can supplement or become an alternative to the web- or computer-based version. Studies have found that paper-based concordance lines offer similar advantages for language pattern analysis (Yilmaz 2017; Jalilifara, Mehrabib, & Mousavinia, 2014). Low-proficiency learners may also benefit from corpus-based learning by using prepared paper materials (Boulton, 2010). Similar to Boulton's argument, there seems to be an agreement that paper-based concordance can be valuable in itself (Huang, 2014; Johns, 1991) or supplemental to provide authentic language data (Breyer, 2006; Frankenberg-Garcia, 2005).



***Lexico-grammatical approach***. The integration of complex syntax and multi-word unit instruction is another recommendation to increase the effectiveness of corpus-based DDL. Previous research examining the reading comprehension of undergraduate medical school students showed that students' processing of texts in a word-by-word manner and their unfamiliarity with complex sentence structures resulted in unsuccessful chunking of ideas into meaningful units (Rasikawati, 2012). The study recommended that reading EAP classrooms need to include the teaching of complex syntax and "extensive practice for the students to chunk ideas at the sentence level" (2012, p. 19). Chan & Liou (2005) similarly suggest that even though the use of concordances scaffolded collocation learning, students with low proficiency levels can perform better with collocation instruction. Ashouri, Armandi, and Rahimi (2014) who studied the impact of corpus-based collocation instruction found that students who learned lexical collocations – chunks of words that often appear together – improved their collocational knowledge better than those learning individual words. The authors claim that corpus-based collocation learning "increases the quantity of learners' mental interaction" and facilitate comprehension improvement (2014, p. 478).

L2 instruction that covers the use of authentic materials and instruction of lexical and syntactic knowledge can be referred to as the lexico-grammatical approach. Lexico-grammar, the unity of lexis and grammar, views lexicon and grammar as two integral parts of a language (Sinclair, 1991). Unlike the traditional grammar approach in language teaching, which consider vocabulary instruction separately from grammar, lexico-grammar instruction perceives learning multi-word sequences or language chunks – words that frequently appear together – in a sentence better facilitates language acquisition. Since learning vocabulary and grammar often takes place simultaneously, instruction of the two domains should be done at the same time (Liu & Jiang, 2009). The acquisition of common lexico-

grammatical patterns is expected to improve students' language ability, including reading as it enables students to process receptive vocabulary faster (Conklin & Schmitt, 2008).

## Implications for Instruction

Insights into the benefits and constraints of corpus-based DDL helps determine appropriate strategies and anticipate problems before adopting DLL in the classroom. The fact that college students have acquired a language before learning L2 thus possessed prior knowledge can be especially beneficial for facilitating students' attention to and analysis of the target language samples. Using corpus-based DDL for this group of students is promising as they can activate relevant schemata to analyze L2 corpus data and make inferences about the L2 patterns found in the data. The authentic learning materials and their relevance to students' study discipline are other contributing factors to the efficacy of the approach. Corpus-based DDL can additionally devise a scheme to improve undergraduate students' comprehension of academic texts through the instruction of multi-word sequences and intensive practice of identifying vocabulary and grammatical structures of specific academic discourse.

Empirical evidence also suggests that corpus-based DDL must take into account the difficulty of analyzing extensive authentic data, especially for the lower proficiency language learners. A paper-based concordance can provide equal benefit for language analysis while allowing scaffolded learning opportunities. This article also argues that EAP reading classrooms need to offer rich exposure to the vocabulary and linguistic patterns of academic prose through the instruction of multi-word sequences. Students' awareness of language chunks can help them cope with complex academic discourse in specific academic disciplines. These multi-word sequences encompass both idiomatic and non-idiomatic expressions (Biber, 2006).



In a nutshell, using a word appropriately depends on more than just knowing the word's definition. Students need to learn how to use the word in relation to other words. These constructs of teaching language in contexts serve as the foundation for the lexico-grammar instruction proposed. Based on this lexical and grammatical connection point of view, learning lexico-grammatical patterns from authentic language data stored in corpora may increase students' analytical thinking, language noticing, and acquisition.

## Corpus-Based DDL Framework to Augment Students' Vocabulary Repertoire

Reviews of inquiry learning theories and previous empirical research examining the use of DDL in L2 classrooms have unraveled essential implications for instruction. In achieving the goal of increasing students' vocabulary size and understanding of word nuances in a Reading EAP program, the student learning experience should comprise students' use of authentic materials and high-order thinking skills; students' engagement in problem-solving; and scaffolded inquiries.

Figure 1 illustrates the framework that the author proposes for enhancing students' vocabulary knowledge. The inquiry-based Reading EAP curriculum should include four curriculum elements – course content, instructional approach, activities, and assessment – within which a framework for enhancing students' vocabulary knowledge is developed. The framework includes key variables of learning structured in a recurring cycle of activities, and which interrelationships of the variables are identifying.

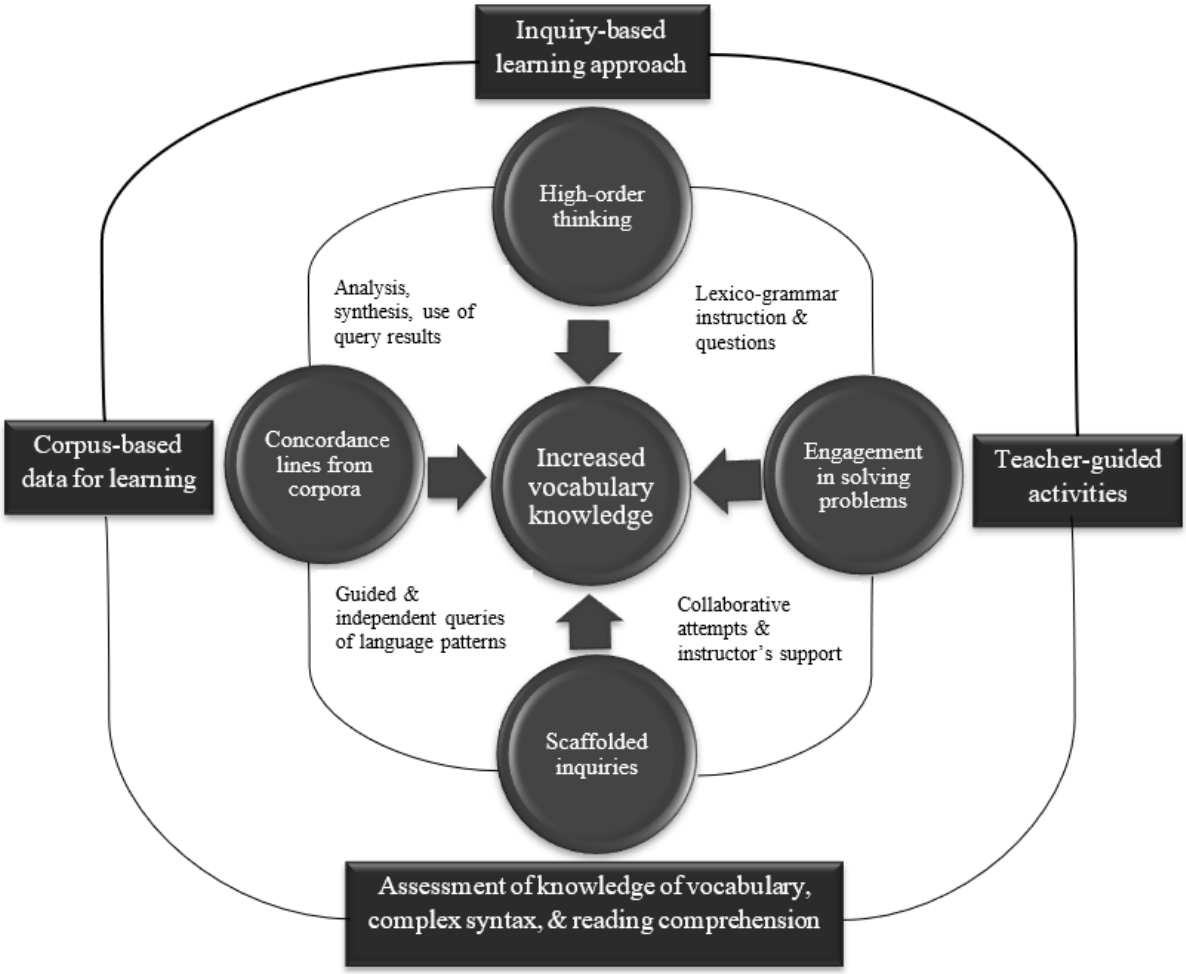


*Figure 1. Framework for augmenting students' vocabulary repertoire*

The cycle begins with instruction on complex syntax and multi-word sequences. The teacher prepares sets of questions to activate background knowledge, model purposeful inquiries, and stimulate high-order thinking. The questions may initially address students' problems as perceived by the teacher but gradually shift to students invented problems and hypotheses as they progress in their learning. Through collaborative work and with the teacher's support, students investigate their problems, collect data from corpus search, discuss how they can analyze the concordance lines, and come up with a synthesis of the lexico-syntactic pattern that emerges from the data. Comparing the patterns that students discover from the texts and the grammatical rules they have learned make their learning more meaningful as they can integrate the information they already know with their findings from corpus consultation activities. As students develop more confidence and better ability to run the corpus-based queries, they learn to solve their lexico-grammatical questions independently, share their findings in a small group or before the class, and practice using their new knowledge in reading and writing. Academic texts for reading activities may include texts in the students' discipline and various teacher-selected texts which type often appear in standardized English proficiency assessments. Including the two types of texts potentially increase students' motivation to read as they are relevant to their study major and need to perform well in a standardized English test, which is often required as a part of the

graduation requirements. Having students use their knowledge of words in writing will help them not only write in response to the reading input but also retain the multi-word sequences learned.

Varied strategies are critical throughout this cycle of activities to facilitate a better understanding of word nuances and syntactical patterns. Paper-based concordance activities can be applied at the early stage of learning before students transition to more substantial web-based corpus data. Sufficient training may likely take several first few meetings to ensure students understand different types of queries to run and results to expect from the queries. The guided inquiries must also include both instruction and modeling of how students can comprehend text by making appropriate chunking of ideas at the sentence level. Ongoing assessments of students' understanding and use of vocabulary in sentences are useful feedback to make any necessary adjustments throughout the program.

## Conclusions and Recommendations for Future Research

Understanding the benefits and constraints of corpus-based DDL helps in determining appropriate strategies and anticipating problems before adopting DLL in the classroom. Corpus-based DDL is a promising approach to enhance undergraduate students' comprehension of academic texts through the use of authentic readings, lexico-grammatical instruction, and analysis of language samples in academic discourse. Reading and analyzing texts that are relevant to their discipline and study interest potentially can increase students' motivation to learn. A corpus-based instructional framework for augmenting students' vocabulary knowledge need to ensure that all the curriculum elements – content, approach, activities, and assessment – are integrated and repeated to achieve the curricular goals, at the same time, promote greater autonomy on the part of the students.

Although many studies have found the use of corpus-based DDL beneficial for L2 students' language development, further research is still needed to provide support for the efficacy of the approach in various contexts. Classroom action research is an alternative to fill the gap between research and instruction in that teachers become more involved in applying the corpus-based DDL approach and evaluating its efficacy. Joint research projects between teachers as the direct users of DDL and the corpus linguistic researchers can establish a stronger connection between empirical evidence and classroom practice. Previous research has documented findings of how corpus linguistics has been used in specific instructional settings. While case studies are valuable channels for gathering detailed accounts of the problem, experimental studies involving large population promises generalization within the given contexts and replication in different settings.



## References

Alfieri, L., Brooks, P.J., Aldrich, N.J., & Tenenbaum, H.R. (2011). Does discovery-based instruction enhance learning? In *Journal of Educational Psychology, 103*(1), pp. 1–18. URL: http://dx.doi.org/10.1037/a0021017

Anderson, R.C., & Freebody, P. (1981). Vocabulary knowledge. In Guthrie, J. (Ed.), *Comprehension and teaching: Research reviews.* Newark, DE: International Reading Association, pp. 77–117.

Anderson, R. C., & Nagy, W. (1992). The vocabulary conundrum. In *American Educator, 16*(4), pp. 14–18, pp. 44–47.

Ashouri, S., Arjmandi, M., & Rahimi, R. (2014). The impact of corpus-based collocation instruction on Iranian EFL learners' collocation learning. *Universal Journal of Educational Research, 2*(6), pp. 470–479. URL: https://doi.org/10.13189/ujer.2014.020604

Bennett, G. R. (2010). *Using corpora in the language learning classroom: Corpus linguistics for teachers*. Ann Arbor: University of Michigan Press.

Biber, D. (2006). *University language: a corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D., & Reppen, R. (2015). Introduction. In Biber, D., & R. Reppen, R. (Eds.), *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press, pp. 2–8.

Blessinger, P., & Carfora, J. M. (2014). Innovative approaches in teaching and learning: An introduction to inquiry-based learning for the arts, humanities, and social sciences. In Blessinger, P., & Carfora, J. M. (Eds.), *Inquiry-Based Learning for the Arts, Humanities, and Social Sciences: A Conceptual and Practical Resource for Educators*. Bingley, UK: Emerald Publishing Limited. Available online at ProQuest Ebook Central. URL: https://ebookcentral-proquest-com.

Boulton, A. (2009). Testing the limits of data-driven learning: language proficiency and training. In *ReCALL, 21*, 37–54. URL: https://doi.org/10.1017/S0958344009000068

Boulton, A. (2010). Learning outcomes from corpus consultation. In M. M. Jaén, F.S. Valverde, & M.C. Pérez (eds.), *Exploring new paths in language pedagogy: Lexis and corpus-based language teaching*. London: Equinox, pp. 129–144. URL: https://files.eric.ed.gov/fulltext/ED544438.pdf

Boulton, A. (2012). What data for data-driven learning? In *The EUROCALL Review: Proceedings of the EUROCALL 2011 Conference* (Vol. 20). Nottingham, UK, pp. 23-27.

Boulton, A. (2017). Corpora in language teaching and learning. In *Language Teaching, 50*(4), pp. 483–506. https://doi.org/10.1017/S0261444817000167

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning 67*(2), pp. 348–393. URL: https://doi.org/10.1111/lang.12224

Breyer, Y. (2006). My Concordancer: Tailor-made software for language learners and teachers. In S. Braun, K. Kohn, & J. Mukherjee (Eds.), *Corpus technology and language pedagogy: New resources, new tools, new methods*. Frankfurt: Peter Lang, pp. 157–176.

British National Corpus (BNC), version 4.4 (BNC XML Edition). (2018). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk.

Bruner, J. (1961). The act of discovery. In *Harvard Educational Review, 31, pp.* 21–32.

Carver, R.P. (1994). Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: Implications for instruction. In *Journal of Reading Behavior 26*(4), pp. 413–437. URL:https://journals.sagepub.com/doi/10.1080/10862969409547861

Chambers, A.(2007). Integrating corpora in language learning and teaching. In *ReCALL, 19*(3), pp. 249–251. URL: https://doi.org/10.1017/S0958344007000134

Chan, T. P., & Liou, H. C. (2005). Effects of web-based concordancing instruction on EFL students' learning of verb–noun collocations. In *Computer Assisted Language Learning, 18*(3), pp. 231–251. URL: https://doi.org/10.1080/09588220500185769

Cheng, W. (2012). *Exploring Corpus Linguistics Language in Action*. New York: Routledge

Clifton, J., & Phillips, D. (2006). Ensuring high surrender value for corporate clients and increasing the authority of the language instructor: The dividends of a data-driven lexical approach to ESP. In *The Journal of Language for International Business, 17*(2), pp. 72–81. URL: http://ezproxy.spu.edu/login?url=https://search-proquest-com.ezproxy.spu.edu/docview/197241495?accountid=220

Coffman, T. (2017). *Inquiry-Based Learning* (3rd ed.). Lanham, MD: Rowman & Littlefield.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? In *Applied Linguistics, 29*(1), pp. 72–89. URL: https://doi.org/10.1093/applin/amm022

Davies, M. (2004–) *BYU-BNC*. (Based on the British National Corpus from Oxford University Press). URL: https://corpus.byu.edu/bnc/ .

Davies, M. (2008–) *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. URL: https://corpus.byu.edu/coca/ .

Dewey, J. (1910). *How we think.* New York: Prometheus Books.

Eadie, G. (2001). *The Impact of ICT on Schools: Classroom Design and Curriculum Delivery: a Study of Schools in Australia, USA, England, and Hong Kong.* Wellington: Winston Churchill Memorial Trust.

Evans, M. (2009). Introduction. In M. Evans (Ed.), *Foreign Language Learning with Digital Technology*. New York: Continuumpp. pp. 1–6.

Farr, F., & Murray, L. (2016). Introduction: Language learning and technology. In F. Farr and L. Murray (Eds.), *The Routledge Handbook of Language Learning and Technology*. New York: Routledge, pp. 1–5.

Frankenberg-Garcia, A. (2012). Learners' use of corpus examples. In *International Journal of Lexicography, 25*(3), pp. 273–296. URL: https://doi.org/10.1093/ijl/ecs011

Frankenberg-Garcia, A. (2005). Pedagogical uses of monolingual and parallel concordances. In *ELT Journal, 59*(3), pp. 189–198. URL: https://doi.org/10.1093/elt/cci038 (retrieved: 2019, September 6).

Geluso, J., & Yamaguchi, A. (2014). Discovering formulaic language through data-driven learning: Student attitudes and efficacy. In *ReCALL*, 26, pp. 225–242. URL: https://doi.org/10.1017/S0958344014000044

Godwin-Jones, R. (2017). Data-informed language learning. Language Learning & Technology, 21(3), pp. 9–27. URL: http://llt.msu.edu/issues/october2017/emerging.pdf/

Hattie, J. (2012). *Visible learning for teachers.* London, NY: Routledge.

Huang, Z. (2014). The effects of paper-based DDL on the acquisition of lexico-grammatical patterns in L2 writing. In *ReCALL, 26*(2), pp. 163–183. URL: https://doi.org/10.1017/S0958344014000020

Hunt, A., & Beglar, D. (2005). A framework for developing EFL reading vocabulary. In *Reading in a Foreign Language 17*(1), pp. 23–59. URL: http://nflrc.hawaii.edu/rfl/April2005/hunt/hunt.pdf/

Jalilifar, A., Mehrabi, K., & Mousavinia, S. R. (2014). The effect of concordance enriched instruction on the vocabulary learning and retention of Iranian EFL learners. In *Procedia – Social and Behavioral Sciences, 98*(6), pp. 742–746. URL: https://doi.org/10.1016/j.sbspro.2014.03.476 (retrieved: 2019, September 7)

Johns, T. (1991). Should you be persuaded – Two samples of data-driven learning materials. In T. Johns & P. King (Eds.), *ELR Journal Vol. 4: Classroom Concordancing,* pp. 1–16.

Johns, T. (1994). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Odlin (Ed.), *Perspectives on Pedagogical Grammar*. New York: Cambridge University Press, pp. 293–313.

Karras, J. N. (2016). The effects of data-driven learning upon vocabulary acquisition for secondary international school students in Vietnam. In *ReCALL, 28*, 166–186. URL: https://doi.org/10.1017/S0958344015000154

Keck, C. (2004). Corpus linguistics and language teaching research: Bridging the gap. *Language Teaching Research, 8*(1), pp. 83–109. URL: https://doi.org/10.1191/1362168804lr135ra

Kennedy, C., & Miceli, T. (2001). An evaluation of intermediate students' approaches to corpus investigation. In *Language Learning & Technology, 5,* pp. 77–90. URL: https://scholarspace.manoa.hawaii.edu/bitstream/10125/44567/1/05_03_kennedy.pdf

Kirschner, P., Sweller, J., & Clark, R. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. In *Educational Psychologist, 41*(2), pp. 75–86. URL: https://doi.org/10.1207/s15326985ep4102_1

Kojic-Sabo, I. & Lightbown, P. M. (1999). Students' approaches to vocabulary learning and their relationship to success. In *The Modern Language Journal*, *83*(2), pp176–192. URL: https://doi.org/10.1111/0026-7902.00014

Kosem, I. (2008). User-friendly corpus tools for language teaching and learning. In A. Frankenberg-Garcia (Ed .), *Proceedings of the 8th teaching and language corpora conferenc*e . Lisbon, Portugal: ISLA. pp. 183–192. URL:

http://anafrankenberg.synthasite.com/resources/TaLCLisbon2008Proceedings.pdf

Kurnia, N. (2003). *Retention of multi-word strings and meaning derivation from L2 reading*. Unpublished doctoral dissertation, Victoria University of Wellington, New Zealand.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size, and reading comprehension. In *Reading in a Foreign Language 22*(1), pp. 15–30. URL: http://nflrc.hawaii.edu/rfl/April2010/articles/laufer.pdf

Levy, B. L. M., Thomas, E. E., Dargo, K., & Rex, L. A. (2013). Examining studies of inquiry-based learning in three fields of education: Sparking generative conversation. In *Journal of Teacher Education, 20*(10), pp. 1–22. URL: https://doi.org/10.1177/0022487113496430

Liu, D., & Jiang, P. (2009). Using a corpus-based lexico-grammatical approach to grammar instruction in EFL and ESL contexts. In *Modern Language Journal, 93*, pp. 61–78. URL: https://doi.org/10.1111/j.1540-4781.2009.00828.x

Liu, D., & Lei, L. (2017). *Using corpora for language learning and teaching*. Alexandria, VA: TESOL International Association.

McEnery, T., & Xiao, R. (2010). What corpora can offer in language teaching and learning. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning*. London & New York: Routledge, pp. 364–380.

Moon, R. (2007). Sinclair, lexicography, and the cobuild project. In *International Journal of Corpus Linguistics, 12*(2), pp. 159–181. URL: https://doi.org/10.1075/ijcl.12.2

Nagy, W. E. (1988). *Teaching Vocabulary to Improve Reading Comprehension*. Urbana, IL: National Council of Teachers of English.

Nagy, W. E., & Townsend, D. (2012). Words as Tools: 'Learning Academic Vocabulary' as Language Acquisition. In *Reading Research Quarterly, 47*(1), pp. 91–108. URL: https://doi.org/10.1002/RRQ.011

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), pp. 59–82.

O'Keeffe, A ., & Farr, F. (2003) Using language corpora in initial teacher training: Pedagogic issues and practical application. In *TESOL Quarterly, 37*, pp. 389–418.

Ornstein, A., & Hunkins, F. (2016). Curriculum design. Historical foundations of curriculum. In A. Ornstein, & F. Hunkins, *Curriculum: Foundations, Principles, and Issues (7th Ed.)*, pp. 181-206. Boston, MA: Pearson/Allyn and Bacon.

O'Sullivan, I. (2007). Enhancing a process-oriented approach to literacy and language learning: The role of corpus consultation literacy. In *ReCALL, 19*(3), pp. 269–286. URL: https://doi.org/10.1017/S095834400700033X

Rasikawati, I. (2012). Medical Students' Comprehension Problems in Reading English Medical Texts. In *Jurnal Kedokteran MEDITEK*, *18*(47): pp. 11–19. URL:

http://ejournal.ukrida.ac.id/ojs/index.php/Ked/issue/view/191

Romer, U. (2008). Corpora and language teaching. In A. Ludeling & M. Kyto (Eds.), *Corpus linguistics: An international handbook* (Vol. 1). Berlin: Mouton de Gruyter. pp. 112–130.

Schmitt, N. (2014). Size and Depth of Vocabulary Knowledge: What the Research Shows. In *Language Learning 64*(4), pp. 913–951. URL: https://doi.org/10.1111/lang.12077

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction*. Cambridge, UK: Cambridge University Press, pp. 3-33.

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning*. Honolulu, HI: University of Hawaii Press, pp. 1–63.



Shaw, E. M. (2011). *Teaching Vocabulary through Data-Driven Learning*. All Theses and Dissertations. 3024. URL: https://scholarsarchive.byu.edu/etd/3024/

Sinclair, J. (2004). New evidence, new priorities, new attitudes. In J. Sinclair (Ed.), *How to use corpora in language teaching*. Amsterdam: John Benjamins, pp. 272-299.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Smart, J. (2014). The role of guided induction in paper-based data-driven learning. In *ReCALL 26*(2), pp. 184–201. URL: https://doi.org/10.1017/S0958344014000081

Snow, C.E., & Uccelli, P. (2009). The challenge of academic language. In D.R. Olson & N. Torrance (Eds.), *The Cambridge handbook of literacy*. New York: Cambridge University Press, pp. 112-133.

Stahl, S. A., & Nagy, W. (2006). *Teaching word meanings.* Mahwah, NJ: Erlbaum.

Szudarski, P. (2018). *Corpus linguistics for vocabulary: A guide for research*. London & New York: Routledge.

Wang, Y. (2016). *The Idiom Principle and L1 Influence: A contrastive learner-corpus study of delexical verb + noun collocations.* Amsterdam & Philadelphia: John Benjamins Publishing Company.

Warren, M. (2016). Introduction to data-driven learning. In F. Farr and L. Murray (Eds.), *The Routledge Handbook of Language Learning and Technology*. New York: Routledge, pp. 337-347.

Yimaz, M. (2017). The effect of data-driven learning on EFL students' acquisition of lexico-grammatical patterns in EFL writing. In *Eurasian Journal of Applied Linguistics 3*(2), pp. 75–88. URL: https://doi.org/10.32601/ejal.460966 . (retrieved: 2019, September 7).

Yoon, H. & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. In *Journal of Second Language Writing, 13*(4), pp. 257–283. URL: https://doi.org/10.1016/j.jslw.2004.06.002

## About the Author

**Ira Rasikawati:** Lecturer, English Department, Universitas Kristen Krida Wacana, Jakarta, Indonesia. Ph.D. in Education Candidate with a Concentration in Literacy, Seattle Pacific University; e-mails: irarras@ukrida.ac.id; rasikawatii@spu.edu

◆ ◆ ◆