Gist Education and Learning Research Journal. ISSN 1692-5777.
No. 27 (July - December, 2023). pp. 69-90.

L2 Learners' Perspectives on Data-Driven Learning for Identifying Properties of Near-Synonymous Words: A Convergent Mixed-Methods Study

Perspectivas de los estudiantes de L2 sobre el aprendizaje basado en datos para identificar propiedades de palabras casi sinónimas: un estudio convergente de métodos mixtos

Dr. Sibel Söğüt[1]

Sinop University

[1] ssogut@sinop.edu.tr

Abstract

This study examines second language (L2) learners' perspectives regarding the affordances and challenges of using the Data-Driven Learning (DDL) to identify the properties of near-synonymous words. Employing a convergent mixed-method design, this study deciphers the perceptions of 40 undergraduate L2 learners majoring in English language teaching. After an initial identification of the learners' vocabulary levels, the experienced benefits and barriers associated with carrying out experiential tasks were elicited via questionnaire data and open-ended survey questions. Descriptive statistics, including means and standard deviations, were revealed and thematic analyses of the responses to the survey questions were documented. The results indicate that completing tasks through the corpus was found to enhance their knowledge of collocations. Integrating corpus tasks into YouGlish (an online practice tool for authentic spoken English in context) was found to increase their awareness of the contextual properties of words. The identification of condensed language exposure, lexical inference, and elicitation of flexible and context-specific patterns were reported to be beneficial. Acknowledging these benefits, gaining familiarity with the corpus interface, encountering limited access to search queries, and analyzing large amounts of concordance lines posed challenges for learners. This research presents the implementation of the DDL supported by experiential learning, contextually rich input, and inductive reasoning tasks in vocabulary learning by further offering instructional implications in L2 contexts.

Keywords: data-driven learning, vocabulary learning, near-synonymous words, L2 learners, learner perspectives.

Resumen

Este estudio examina las perspectivas de los estudiantes de L2 con respecto a las posibilidades y desafíos del uso del aprendizaje basado en datos (DDL) para identificar las propiedades de palabras casi sinónimas. Empleando un diseño de método mixto convergente, este estudio descifra las percepciones de 40 estudiantes universitarios de L2 con especialización en enseñanza del idioma inglés. Después de una identificación inicial de los niveles de vocabulario de los estudiantes, los beneficios experimentados y las barreras asociadas con la realización de tareas experienciales se obtuvieron a través de datos de cuestionarios y preguntas de encuesta abiertas. Se revelaron estadísticas descriptivas, incluidas medias y desviaciones estándar, y se documentaron análisis temáticos de las respuestas a las preguntas de la encuesta. Los resultados indican que completar tareas a través del corpus mejora su conocimiento de las colocaciones y aumenta su conciencia de las propiedades contextuales de las palabras con la intersección del corpus y el YouGlish. Se informó que la identificación de la exposición al lenguaje condensado, la inferencia léxica y la obtención de patrones flexibles y específicos del contexto eran beneficiosas. Reconocer estos beneficios, familiarizarse con la interfaz del corpus, encontrar acceso limitado a consultas de búsqueda y analizar grandes cantidades de líneas de concordancia planteó desafíos para los estudiantes. Esta investigación avanza en la implementación de DDL respaldada por el aprendizaje por descubrimiento, aportes contextualmente ricos y tareas de razonamiento inductivo en el aprendizaje de vocabulario al ofrecer implicaciones instructivas en contextos de L2.

Palabras clave: aprendizaje basado en datos, aprendizaje de vocabulario, palabras casi sinónimas, estudiantes de L2, perspectivas del estudiante.

Resumo

Este estudo examina as perspectivas dos alunos de L2 em relação às possibilidades e desafios do uso da aprendizagem baseada em dados (DDL) para identificar as propriedades de palavras quase sinônimas. Utilizando um desenho de método misto convergente, este estudo analisa as percepções de 40 estudantes universitários de L2, com especialização em ensino de língua inglesa. Após uma identificação inicial dos níveis de vocabulário dos alunos, os benefícios e as barreiras experimentadas associados à realização de tarefas experienciais foram obtidos por meio de dados de questionários e perguntas abertas de pesquisa. Foram reveladas estatísticas descritivas, incluindo médias e desvios-padrão, e documentadas análises temáticas das respostas às perguntas da pesquisa. Os resultados indicam que a realização de tarefas por meio do corpus melhora o conhecimento das colocações e aumenta a conscientização sobre as propriedades contextuais das palavras com a interseção do corpus e do YouGlish. Foi relatado que a identificação da exposição à linguagem condensada, a inferência lexical e a obtenção de padrões flexíveis e específicos de contexto foram benefícios observados. Reconhecer esses benefícios, familiarizar-se com a interface do corpus, encontrar acesso limitado a consultas de pesquisa e analisar grandes quantidades de linhas de concordância apresentaram desafios para os alunos. Esta pesquisa avança na implementação de DDL apoiada pela aprendizagem por descoberta, com entradas contextualmente ricas e tarefas de raciocínio indutivo no aprendizado de vocabulário, oferecendo implicações instrutivas em contextos de L2.

Palavras-chave: aprendizagem baseada em dados, aprendizado de vocabulário, palavras quase sinônimas, estudantes de L2, perspectivas dos estudantes

Recent developments in second language (L2) acquisition research have yielded growing interest in vocabulary teaching. Near-synonyms have meaning differences in terms of their denotational variations (i.e. propositional, fuzzy, and other peripheral aspects), stylistic variations (i.e., dialect and register), expressive variations (i.e., emotive and attitudinal aspects), and structural variations (i.e., collocational, selectional, and syntactic variations) (Cruse, 1986). To date, several studies have shown that existing bilingual dictionaries are not always helpful in conveying subtle differences among near-synonyms, as they highlight denotation rather than usage (Xiao & McEnery, 2006).

Such features of near-synonyms pose a stumbling block to L2 learners' lexical choices. The demanding nature of the learning properties of near-synonyms has a significant influence on learners' affective factors and their overall performance in L2 acquisition. It can be difficult even for native speakers to identify the differences between near-synonyms well enough to use them and "Choosing the wrong word can convey an unwanted implication" (Edmonds & Hirst, 2002, p. 108). Near-synonymy inherently affects the structure of lexical knowledge (Edmonds & Hirst, 2002, p. 106), and learners need to observe repeated patterns and meanings to identify differences originating from collocational behavior and semantic prosody (Xiao & McEnery, 2006). Considering these features, figuring out the differences between near-synonyms and making appropriate lexical choices when learning new vocabulary can be a particularly challenging endeavor (Lin & Chung, 2021). L2 learners' mastery of near-synonymous words may be enhanced by providing authentic contexts and integrating corpus tools in L2 learning processes.

To address these challenging aspects, DDL has emerged as a promising pedagogical endeavor that enables access to exploratory activities for the acquisition of vocabulary knowledge. This technique provides space for learners to learn by exploring and analyzing language data from a corpus (Johns, 1986). It also offers pedagogical benefits by introducing new phraseology to young learners (Szudarski, 2019). It also enables learners to engage in authentic concordance lines by promoting their autonomy and awareness to successfully discover pattern regularities (Szudarski, 2022). These tasks serve to help learners overcome different types of vocabulary errors and improve their academic writing quality (Alsehibany & Abdelhalim, 2023).

This learner-centered technique provides space for hands-on practices, along with the discovery of learning experiences. It also provides a platform for the application of critical thinking skills, noticing, gaining awareness about language samples, creating and testing hypotheses, acquaintance with linguistic variation, and data analysis skills (Pérez-Paredes et al., 2019). Previous empirical studies have established the basis and connections between corpus-driven tasks and L2 skill acquisition. Much research has documented learners' corpus use behaviour and their perceptions of the strengths and weaknesses of corpora as a second language writing tool (Yoon & Hirvela, 2004;

Flowerdew, 2010). Examining the effects of integrating corpus and contextualized lexico-grammar in L2 teaching, Liu and Jiang (2009) documented that analyzing concordance data to identify lexico-grammatical usage rules and patterns is the greatest challenge for learner. A meta-analysis revealed that the level of proficiency in L2 and various features of the corpus use (i.e., types of interaction, types of corpora, training, and duration) affect the extent to which corpus use enhances L2 vocabulary acquisition (Lee et al., 2019).

The existing literature on the pedagogical applications of DDL is extensive. An overview of the prevailing discourse regarding the use of corpus tools in enhancing L2 learners' vocabulary acquisition focuses on pedagogical benefits. Previous research has documented a meta-analysis of DDL in English as a Foreign Language (EFL) classroom in the Japanese context and revealed that learners exhibited substantial learning gains for acquiring the properties of lexical items, expanding their repertoires of grammar and formulaic sequences (Mizumoto & Chujo, 2015). Previous seminal research has established that the use of corpora enhances language learning and teaching through authentic language input (Gilquin, 2022; Lei & Liu, 2018). Corpora also expose learners to contextualized language samples and quantitative information (Gilquin, 2022). Learners take on the role of language detectives or researchers, exploring authentic examples of the target language through corpus-based tasks (Geluso & Yamaguchi, 2011). It provides access to contextual analysis of numerous samples of authentic language use (Sevilan, 2023).

Considering these features, previous research has established that students believe DDL is a useful and effective tool in the classroom (Geluso & Yamaguchi, 2014) as it enhances their critical understanding of grammar and discovery learning skills (Liu & Jiang, 2009). DDL tasks serve the potential for learners to establish connections between these patterns and their respective meanings even at lower levels if they are provided with carefully selected patterns presented in a contextually rich format (O'Keeffe, 2023). Such practices enable learners to carry out hands-on concordancing and foster critical reading skills (Yang & Mei, 2024). Accordingly, Leńko-Szymańska (2022) argues that corpus-related pedagogical skills, which entail technical and corpus-analytical skills, should be integrated into language teacher training.

Along with the previously reported pedagogical gains, Boulton (2010) lists the limitations of corpus use in language learning, including a) new material (e.g., keywords in context format), b) technology (e.g., concordancer), and c) learning approaches (e.g., inductive learning). The time-consuming nature of DDL and the difficulty in interpreting the results of corpus investigations have also been highlighted in the literature (Yoon & Hirvela, 2004). The potential of drawing wrong inferences and 'fake discovery' (O'Keeffe, 2023), loaded or insufficient output of the search queries, teachers' lack of knowledge, and awareness of corpus applications in language classes (Gilquin & Granger, 2022) are additional reported limitations of the DDL.

The nature of near-synonymous words poses challenges to L2 learners. Several corpus-based analyses have documented the properties of near-synonymous words in English (Lin & Chung, 2021; Song, 2021). These studies demonstrate that near-synonymous words are not used interchangeably (Edmonds & Hirst, 2002), are not fully intersubstitutable (Song, 2021) and operate in different contexts (Xiao & McEnery, 2006) because of their semantic, syntactic, and pragmatic properties. Although near-synonyms have distinct semantic profiles, dictionaries present them as interchangeable in different contexts, and this presentation may guide L2 learners to assume contextual interchangeability (Alanazi, 2022).

Particular challenges have also been reported in the acquisition of near-synonymous words. Previous research has revealed that L2 learners' inappropriate use of near-synonymous words may stem from several factors including interference of L1, inadequate descriptions of these words in dictionaries, and insufficient instructional focus on the subtle semantic differences among synonyms (Liu, 2018). To date, several studies have shown that existing bilingual dictionaries are not always helpful in conveying subtle differences among near-synonyms as they highlight denotation rather than usage (Xiao & McEnery, 2006). A significant argument proposes that native speaker introspection is no longer considered the sole, reliable source of insight into language structure and is used to document these properties and differences (Gabrielatos, 2005).

Overall, the existing arguments uncover gaps and notably scarce literature regarding the use of data-driven learning to practice the properties of near-synonymous words in a teacher training context. By exploring the attitudes of pre-service teachers and eliciting their perceptions of and practices regarding the DDL, practical implementations derived from experienced barriers and benefits can offer insights into integration of the DDL into teacher training. Further, exploration of learning several properties of near-synonymous words can serve to better understand potential ways of integrating the DDL into vocabulary acquisition in L2 contexts. Drawing upon this highlighted need and previously documented pedagogical benefits, this study frames the investigation of L2 learners' experiences in a teacher-training context as an underexplored area. This study aims to contribute to this growing area of interest by exploring L2 learners' experiences of conducting DDL experiential learning tasks to decipher pedagogical benefits and potential drawbacks. This study was motivated by the pedagogical affordances of the DDL approach, and the complexity and challenging nature of the properties of near-synonymous words in English. Studies on the topic focus more on the benefits and limitations originating from the tool and instructional design. In contrast to previous research, this study offers a fresh perspective and addresses L2 learners' experiences in an EFL teacher education context, where they have limited technological tools and digital sources due to the existing digital divide in their setting.

A critical view of the aforementioned studies shows that there is a tendency to document positive results on the substantial learning gains of corpora use and the widespread implementation of corpus-based tasks. This study examines the emerging role of the DDL in deciphering a composite picture of lived experiences concerning experiential reflections and potential barriers. This study revisits the need to document a cluster of evolving learning gains and examines L2 learners' perspectives on the opportunities and challenges of using a corpus-based data-driven learning approach to practice the properties of near-synonymous words in a vocabulary course. This study places the DDL at the center of the course syllabus to enhance pre-service teachers' corpus literacy skills. A novel contribution of the current study is the documentation of affordances of an array of functions with the intersection between the COCA and Youglish, and revisiting this landscape from the perspective of prospective English language teachers. This study outlines a corpus-based vocabulary teaching course with the aim of providing authentic language input; disrupting heavy reliance on textbooks; actively engaging learners in their discovery learning processes; and conducting an in-depth analysis of the properties of near-synonymous words. This study aims to uncover learner perspectives and is driven by the following questions:

1. What are the pedagogical benefits and potential drawbacks of incorporating a DDL approach for teaching near-synonyms in a teacher education context with limited technological tools and digital sources?

2. How do L2 learners experience and perceive the opportunities and challenges of conducting DDL experiential learning tasks and how does this approach enhance their corpus literacy skills and engagement in their discovery learning processes?

Method

Research Setting and Participants

This study was conducted within the scope of a vocabulary course delivered at a Turkish state university. The participants were 40 undergraduate pre-service teachers majoring in the English Language Teaching (ELT) department. The learners were administered an institutional English proficiency test at the beginning of the semester, and were also involved in a two-semester preparatory program to gain mastery over skill-based courses before enrolling in an undergraduate degree in the ELT department. They took writing, speaking, listening, vocabulary, and reading skills courses, and English was the medium of instruction in these courses. Their English proficiency level was B1 as described in the Common European Framework of Reference for Languages.

Procedures

This study was conducted as part of the Vocabulary Course. As part of the ethical guidelines, informed consent was received from each participant. The participants were informed about the purpose of the study, and the anonymity of their responses was ensured by eliminating any identifying information in the data-gathering tools. Initially, the Vocabulary Levels Test (VLT) by Schmitt et al. (2001) was conducted at the beginning of the term to gain an understanding of learners' lexicons and identify their needs. This test provides an estimate of vocabulary size for L2 learners of general and academic English (Schmitt et. al., 2001), with a focus on the most frequently used words in English. The test consists of words required in basic everyday oral communication (2000-word level), reading authentic texts (3000-word level), inferring the meanings of novel words from context and understanding the communicative content (5000-word level), and having knowledge of the sub-technical vocabulary occurring across a range of academic disciplines (10000) (Schmitt et al., 2001). Academic Word Level (AWL) provides an estimate of the size of learners' academic vocabulary (Schmitt et al., 2001). The AWL was placed between 3000 and 5000 sections, as the placement of this section is flexible based on the demands of each testing situation (Schmitt et al., 2001). Considering the need to obtain an estimate of vocabulary size, the pedagogical needs of the learners in terms of the properties of words, and their tendency to cope with authentic language input, this benchmark was utilized. Learners' levels are presented in Table 1.

Table 1. Vocabulary Levels (VL) of the Learners

VL	2000	3000	AWL	5000	Total
N	11	15	10	4	40
%	27,5	37,5	25	10	100

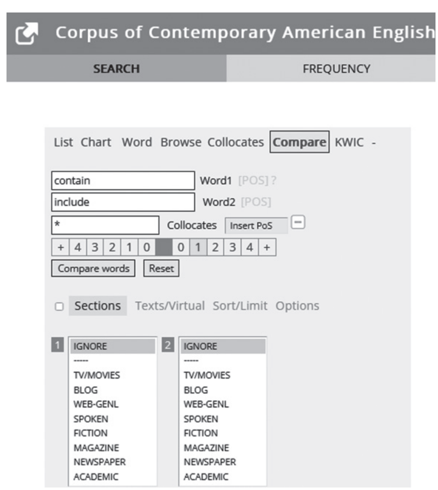
The results of the VLT show that the word levels of most learners were at the 3000-word level, followed by 2000 and AWL.

The Corpus of Contemporary American English (COCA) (Davies, 2008) was used as a tool to integrate DDL into the vocabulary course. COCA is a genre-balanced corpus containing more than one billion words of text from various genres, such as spoken, fiction, popular magazines, newspapers, academic texts, TV and movie subtitles, blogs, and other web pages. The design of the tasks included in the course content was derived from an array of suggestions and descriptions of corpus-driven pedagogical materials provided by Gabrielatos (2005). Two sessions of 45 minutes of corpus training and subsequent administration of hands-on practices were carried out with the employment of corpus-based tasks over the course of 14 weeks throughout the spring semester.

Prior to the course, learners were introduced to the course content, syllabus, requirements, and objectives. From an emic perspective, the learners' prior corpus use was initially elucidated. It was figured out that the learners had no prior experience or familiarity with the use of a corpus in their language learning processes. As the comprehension of concordancing would be difficult without teacher instruction (Boulton & Cobb, 2017), the learners were trained to conduct search queries with different functions (i.e. distribution across years, registers, collocational patterns, etc.) and were familiarized with the interface and an array of functions of the corpus throughout the course. The course content was designed to decipher the multifaceted properties of target words. DDL tasks were employed in the course, and learners were assigned to both in-class and out-of-class tasks to figure out properties of near-synonymous words by administering hands-on practices in the corpus.

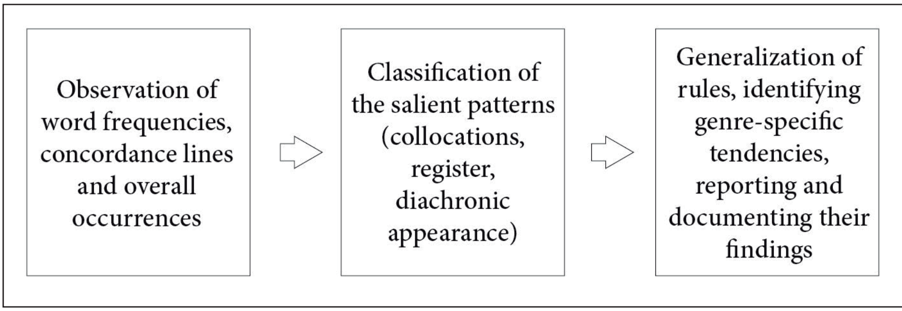
A blended learning approach was used throughout the course, with the integration of researcher-prepared tutorials, and discussion platforms set on Canvas to enable learners to discuss their findings outside the classroom. The in-class practice sessions included teacher-directed corpus-driven tasks, guided corpus queries, and learner-centered discovery-learning tasks. Figure 1 illustrates a sample guided corpus query provided to the learners.

Figure 1. A Sample Guided Corpus Query Using the COCA



Based on this pedagogical position of DDL, learners prepared reports presenting findings generated by the corpus and discussed them in classroom sessions. The three stages of inductive reasoning proposed by Carter and McCarthy (2004) were implemented in the classroom, as shown in Figure 2.

Figure 2. The Stages of Inductive Reasoning Proposed by Carter & McCarthy (2004)



Following these stages, in-class discussions and feedback sessions were conducted based on the learners' findings, and extended concordance displays for the target words were examined in the classroom. The main parts of a word examined by the learners were parts of speech, synonyms and antonyms, collocational patterns, register information, genre-specific tendencies, formality level, diachronic changes, grammatical behavior of the words, and example sentences. In addition to the corpus, YouGlish was used to provide complementary support as a YouTube-based pronunciation dictionary. This tool was developed to meet the need for authentic pronunciation input, which allows quick and easy access to "YouTube-sourced pronunciation samples" (Barhen, 2019, p. 2). This tool was used to guide learners in examining dictionary definitions, phonemic descriptions, and the pronunciation of the target words. Sample tasks assigned to the learners are exemplified in Appendix A.

Research Design and Instruments

Drawing on the mixed-methods research paradigm, the current study employed a convergent mixed-methods design in which qualitative and quantitative data were collected in parallel, analyzed separately, and then merged (Creswell, 2014). This research design provided a complementary perspective on the benefits and barriers of using corpus tools in vocabulary learning. The quantitative data of this study came from learner responses to a 6-point Likert scale questionnaire from Geluso and Yamaguchi's (2014) study, which combined the items from two published studies on using corpora in L2 writing (Yoon & Hirvela, 2004; Liu & Jiang, 2009). After obtaining consent from the learners, the data were collected at the end of the term, and the learners were asked to indicate their degree of agreement with the items. The instrument consisted of statements about the difficulty in using corpora, the positive impact of using corpora, the effectiveness of presentation and delivery of coursework, attitudes, and beliefs about data-driven learning and its potential. After the administration of the questionnaire, the learners were asked to answer the open-ended questions provided in Appendix

B. Follow-up open-ended questions were semi-structured, with lead questions based on the survey results. The first set of questions elicited learners' experiences of the challenges in using the corpus. The second question sought the learners' perceptions of the most useful and valuable things they learned. The last examined the changes in their perspectives and future orientation about the integration of corpus-driven tasks in their future language classes as pre-service English teachers.

Data Collection and Analysis

A total of 40 L2 learners responded to the questionnaires, while 16 learners responded to follow-up open-ended questions on a voluntary basis. The data collection procedure was conducted once the terms ended and instructor assessments, evaluations, and reflections were finalized. The qualitative and quantitative data collection process was carried out in parallel stages. Initially, the participants were assured of confidentiality through anonymous responses to the study. In the first phase, the participants responded to the questionnaire items and shared their perspectives on the corpus integration. Then, they were asked follow-up open-ended questions to delve into their experiences and perceptions. The findings from both quantitative and qualitative data were compared and combined to provide a comprehensive picture of learner experiences.

The responses revealed from the questionnaire and open-ended questions were analyzed separately. For the quantitative data analysis, means and standard deviations were calculated based on the responses of the participants. This analysis revealed the distribution of learner responses to each item for the identification of common patterns. As for the qualitative data analysis, a thematic analysis was used to identify, analyze, and interpret themes emerging from the learner responses to open-ended questions. At this stage, the whole data was coded, and repetitive segments were assigned codes. Then, codes were grouped into broader themes representing the prominent perspectives shared by the learners. The elicited codes and themes identified based on the responses to the open-ended questions are provided in Appendix C. The key patterns identified in the qualitative data served to support quantitative findings. The results revealed from both qualitative and quantitative data analyses were merged. The results revealed from both sources enabled the documentation of a composite picture of L2 learners' perspectives on experiential learning tasks.

Findings

L2 Learners' Perceptions about the Benefits of Using Corpus

First, the learners reflected on the benefits of the DDL; the overall findings are presented in Table 2. The table shows that learners believed the corpus to be helpful for language learning. Most learners reported that corpus use was most helpful for learning the usage of vocabulary and phrases, meaning of vocabulary, enhancing English reading and writing skills, and increasing their confidence in English vocabulary. A slight decrease was observed in the perceived usefulness of corpus use over a dictionary. A particularly counterintuitive finding was that the scores for learning grammar and improving academic writing ability were relatively low, which should be further elaborated in future investigations.

Table 2. Benefits of Corpus Use (N=40)

Category	Agree (%)	Disagree (%)	M	SD
More helpful than a dictionary for my English vocabulary.	75	25	4.03	1.07
Learning the meaning of vocabulary	90	10	5.00	1.24
Learning the usage of vocabulary	100	0	5.53	0.59
Learning the usage of phrases	100	0	5.50	0.67
Learning grammar	60	40	3.65	1.47
Improved English reading skills	95	5	4.10	1.15
Improved English writing skills	95	5	4.45	1.21
Improved English academic writing ability	45	55	5.00	0.84
Increased my confidence about English vocabulary	85	15	4.90	1.05

1: strongly disagree, 2: disagree, 3: somewhat disagree, 4: somewhat agree, 5: agree, 6: strongly agree

With the overall picture of the learners' perceptions about the benefits of using the corpus in their vocabulary learning practices, this study revealed findings related to three associative aspects in the qualitative data analysis: (a) benefits of vocabulary learning, (b) benefits about other language skills, and (c) affective benefits. Considering the usefulness of corpus-driven activities in vocabulary learning, the learners drew attention to two significant and connected themes: their increased awareness of word properties and their overall reporting of enhanced language competencies. They elaborated their enhanced understanding of word properties in relation to a better understanding of the multiple meanings of words, word associations, parts of speech, frequency counts, collocational patterns, contextual features of words, diachronic changes, genre-specific information, subtle differences between near-synonymous words, and guessing the meaning of unknown words in context. In relation to their overall reports of enhanced language competencies, they reported expanded vocabulary knowledge, awareness of the harmony among words, significance of vocabulary items in English, and contextual features of the words in English. The participants demonstrated that corpus-driven tasks provided benefits in the mastery of other language skills by facilitating writing skills (form sentences, using collocations, and formality level), improving communication and self-expression skills, improving speaking skills, using language strategies, enhancing language analysis skills, and exposure to different usages and sources of language. The benefits of corpus-based vocabulary learning activities were further evidenced by the fact that the learners highlighted affective aspects with a focus on increased elements of curiosity, enhanced excitement to play with words, happiness in becoming familiar with a huge database, and enhanced motivation and enjoyment in searching for learning. Overall, these findings suggest that corpus-based deductive learning activities enhance and enrich learners' vocabulary learning experiences by boosting their interests.

L2 Learners' Challenges in the Use of Corpus

After an in-depth understanding of the benefits, the first set of questions also unpacked the L2 learners' challenges in the use of the corpus. An intriguing look at the concerns and difficulties with respect to corpus use revealed that the learners' reactions to the challenges in corpus use were clustered in a 2.40-3.98 score range, indicating difficulties and obstacles. Table 3 shows that the amount of time and effort necessary to analyze language expressions, limited access to computers or the Internet, unfamiliar vocabulary in concordance lines, and performing search techniques were the main difficulties highlighted by the learners.

Table 3. L2 Learners' Difficulties in Corpus Use (N=40)

Category	Agree (%)	Disagree (%)	M	SD
Limited access to computer/Internet	45	55	3.38	1.75
The speed of Internet connection	25	75	2.20	1.22
Time and effort spent on analyzing the data	67,5	32,5	3.98	1.54
Unfamiliar vocabulary on concordance/ collocate output	47,5	52,5	3.35	1.09
Cut-off sentences in concordance output	30	70	2.90	1.27
Too many sentences in concordance output	30	70	2.95	1.26
The limited number of sentences in concordance output	45	55	2.95	1.30
Analyzing concordance output	20	80	2.50	0.98
Analyzing the collocate output	25	75	2.45	1.13
Performing the search technique	50	50	3.25	1.59
Too difficult real texts	15	85	2.40	0.98

1: strongly disagree, 2: disagree, 3: somewhat disagree, 4: somewhat agree, 5: agree, 6: strongly agree

An inspection of the findings of the qualitative data analysis revealed that the challenges were framed around two emerging themes: challenges originating from the corpus interface and language content. The prominent challenges were limited usage, a need for a premium account, upgraded options, confusion about the interface, analyzing huge amounts of concordance lines, complicated functions, searching techniques, showing unrelated results, and the need to make manual checks. Understanding the genre-specific language content of registers (e.g., news, academic), understanding concordance lines, a limited number of search queries, unfamiliar corpus interface in the initial stages, complex sentences in the corpus, and searching techniques were the main challenges regarding the language content of the corpus.

L2 Learners' Overall Evaluations of the Use of Corpus

The second set of questions uncovered learners' overall evaluations of the use of the corpus in learning the properties of near-synonymous words. As shown in Table 4, the learners shared positive attitudes and feelings toward these activities in the classroom.

Table 4. L2 Learners' Overall Evaluations of the Use of Corpus (N=40)

Category	Agree (%)	Disagree (%)	M	SD
The search technique was easy to learn	70	30	4.20	1.20
Hands-on practice was useful	95	5	4.83	0.84
Use the corpus by own choice	45	55	3.40	1.44
Understand the purpose of using the corpus	95	5	5.30	0.85
Get the information that I need in the corpus	100	0	5.05	0.81
Learn more, like more	80	20	4.45	1.17
Use corpus in the future	95	5	5.35	0.92
Earlier familiarity would be better	85	15	4.45	1.30
A useful resource for English vocabulary	95	5	5.25	1.00
Should be introduced in all vocabulary courses	90	10	5.35	1.07
Should be taught in English classes	90	10	5.10	1.27

1: strongly disagree, 2: disagree, 3: somewhat disagree, 4: somewhat agree, 5: agree, 6: strongly agree

Except for the items regarding their use of the corpus by their own choice, which may be due to a course requirement, all the items revealed strong agreement among the learners. The learners considered the corpus a useful tool for learning and practicing English vocabulary. The prominent positive evaluations revealed from the descriptive analysis were the availability of the information they needed to learn words, the usefulness of hands-on practices, the relevance of the purpose, the usefulness for other courses, and the willingness for further and future uses. Building on this descriptive analysis, an in-depth exploration of learner evaluations through the qualitative data revealed that they would use future corpus studies to better understand the language, to use the corpus in enhancing all language skills, to make better progress in language, to figure out generalizations about language, to check for confusing words, and to discover the difference between near-synonymous words. The learners also highlighted the need to receive more training for using the corpus in academic and daily studies and to analyze the language and links between words. Regarding their overall opinions, most learners reported that the corpus was a reliable source and a fun activity for learning the language. The learners also shared their potential goals for future corpus-based activities in their prospective classes. They highlighted that they would like to introduce the COCA to their learners, use it in the same way as the teacher, and plan to use it to teach words in reading texts.

Discussion, Conclusion, and Suggestions

The results presented in this study enable us to understand L2 learners' perspectives on the benefits and challenges of using a discovery learning corpus-based approach to practice the properties of near-synonymous words through DDL tasks. This study presents learners' experiences of L2 learners' experiential corpus-based tasks with a specific focus on exploratory learning. Based on these results, the overall conclusion, related discussion points, limitations, and future research directions are presented.

As a conclusion for the first research question, the salient benefits of corpus-driven vocabulary learning tasks are learning the meanings and collocations of near-synonymous words in context, improving reading and writing skills, and facilitating learners' confidence in English vocabulary. The most prominent finding to emerge from this part is that corpus tools are perceived as useful in terms of making the learners remember what they work to find out, providing authentic language input, evoking the element of curiosity, raising lexical awareness, enhancing the understanding of contextual features of the words, actively engaging the learners in the language learning process, and fostering autonomous learning experiences. These benefits are supported through the facilities of the corpus in emphasizing the properties of words such as their register, part of speech, and morphological processes; teaching vocabulary in context with a focus on collocations; surrounding elements in concordance lines; presenting diachronic information and genre-specific tendencies of words; formality levels of the words; and subtle differences between near-synonymous words. Because learning the properties of near-synonymous words is a pedagogically challenging task, the learners' aforementioned benefits may be related to the nature of DDL and the exploratory and discovery-oriented vocabulary learning experience provided to them. More specifically, as documented in previous research, DDL embraces concepts of learner autonomy, induction, exemplar-based learning, and constructivism and it enables learners to autonomously explore linguistic patterns, instead of being provided with predigested rules (Boulton & Cobb, 2017). These multilayered benefits have a facilitative effect on learners' retention of near-synonymous words. Additionally, DDL proves itself to be an effective language learning method as it changes the very nature of the L2 classroom (Karras, 2016) by enabling active participation, discovery learning, willingness and motivation to do research, and inductive reasoning by identifying different properties of words. These benefits would have a sustainable impact on learners' future teaching practices, as they provide corpus training to pre-service English language teachers. Empowerment of the benefits of corpus use and provision of corpus training in initial teacher training would enable further integration of corpora into classroom practice (Leńko-Szymańska, 2022; Szudarski, 2022; Zareva, 2017).

Another notable finding is that spending too much time and effort on analyzing the data and concordance lines, limited queries, and confusion about the corpus interface posed difficulties to the learners. The learners highlighted certain challenges concerning the guidance of corpus use, along with emphasizing barriers to the corpus.

Given the multilayered notion of technology-driven challenges, much of the discourse on the corpus tool is framed around the corpus interface. One of the challenges is learners' unfamiliarity with the use of the corpus, its features, and its functions as a tool. L2 learners, who use the corpus for the first time, need the teacher's support as they expect proper and prolonged teacher assistance to get the maximum benefit from it (Sinha, 2021). To achieve this, providing technical support guiding learners to interact using a corpus would promote their autonomous discovery of the language. Once learners gain familiarity with the corpus interface, they can move on to more divergent or autonomous tasks (Geluso & Yamaguchi, 2014). For this reason, there is also an overall need for substantial and specialized training in digital literacy (Pérez-Paredes et al., 2019). More specifically, initial corpus training is compulsory to facilitate learners' corpus literacy (Yang & Mei, 2024). This argument is echoed by Selivan (2023), who argues that learners can be encouraged to establish an initial form-meaning link and move on to more contextual aspects of vocabulary practice through concordancing. In this regard, the teacher plays a crucial role as it contributes to learners' positive attitudes toward using the corpus (Yoon & Hirvela, 2004). Access to technological tools is also reported as a barrier to their effective access to corpus data. Although Pérez-Paredes (2019) noted that access to technology was not identified as an impeding factor, this research reveals a different finding by documenting digital divide-related drawbacks. Overall, additional training and assistance would help learners overcome technical impediments and further develop corpus analytical and literacy skills.

Moving on to the second research question, this study documents that L2 learners have a positive attitude toward corpus-driven vocabulary teaching tasks. The findings in relation to their attitudes contribute in several ways to our understanding of corpus-based vocabulary teaching and provide a basis for using this approach to teach the properties of near-synonymous words. In contrast to the view that low-proficiency learners may not benefit from corpus use, this study notes numerous benefits without downplaying these challenges. To overcome potential challenges, this study suggests that there is a need to enhance learners' mastery of corpus consultation skills through teachers' mediation and support. The key to managing synonyms for L2 learners is to increase exposure to these words and present their salient collocations in meaningful contexts (Liu & Zhong, 2016). This study also revisits the need to enhance corpus literacy skills and integrate corpus tools into teacher training contexts. This integration may be achieved by using corpora to select relevant vocabulary, developing language syllabi and pedagogical materials, and using corpus data as a teaching technique (Szudarski, 2022).

Another notable suggestion of this study is the emerging need to enable L2 learners' exposure and engagement with concordance lines, which provides rich context information (Lin & Chung, 2021) to enhance and enrich their understanding of word properties. Learners use inference skills and verify their inferences by using visual expressions or providing evidence from concordance (Yang & Mei, 2024).

For this reason, it is crucial that the concordance lines provided in the corpus are comprehensible to learners and offer enough contextual clues to facilitate their exploration and understanding of target lexical items during their linguistic investigations (Lee et al., 2019). A particularly interesting observation that results from the analysis of open-ended questions is that they provide deep insight into L2 learners' attitudes toward the challenges and perspectives. Addressing the challenges in dealing with the subtle differences between near-synonymous words, providing continuous encouragement, and boosting their motivation may help learners overcome the reported barriers. The current results are significant in that the learners have surface-level challenges (i.e., corpus software, time), and these challenges can be overcome through familiarity with these learning experiences, hands-on practices, teacher modeling, and fostering autonomous language learning processes. Teacher mediation, the provision of a rich multimodal context (O'Keeffe, 2023) and transforming 'data-driven learning' into 'data-driven use' for autonomous learning are suggested (Gilquin & Granger, 2022).

Considering these findings of the study, this study has some limitations. Given the idiosyncratic nature of each educational context and learner characteristics, this study is limited in terms of a specific sample size which can potentially lead to a lack of generalizability of the benefits and challenges. The collection of additional data through dairies may provide another complementary perspective for elaborating on the findings. Further, a longitudinal study may expand and enhance our understanding of the long-term effects and dynamics of data-driven applications for language education. For future research, analyzing teachers' attitudes toward using corpora in their classes would present a composite picture for better applications. Additionally, the consequences of corpus-driven materials and tools observed in different local settings may uncover dynamic and effective variables through case studies. A follow-up study could examine learners' vocabulary levels after the implementation of data-driven learning tasks. Another study could investigate the long-lasting impacts of corpus-based training on the participating learners by uncovering their integration of corpora into their teaching practices.

References

Alanazi, Z. (2022). Corpus-based analysis of near-synonymous verbs. *Asian-Pacific Journal of Second and Foreign Language Education*, *7*(1), 1-25. https://doi.org/10.1186/s40862-022-00138-5

Alsehibany, R. A., & Abdelhalim, S. M. (2023). Overcoming academic vocabulary errors through online corpus consultation: the case of Saudi English majors. *Computer Assisted Language Learning*, 1-27. https://doi.org/10.1080/09588221.2023.2249503

Barhen, D. (2019). Youglish. *The Electronic Journal for English as a Second Language, 23*(2), 1-10.

Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language Learning, 60*(3), 534-572. https://doi.org/10.1111/j.1467-9922.2010.00566.x

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

Carter, R., & McCarthy, M. (2004). Talking, creating: interactional language, creativity, and context. *Applied Linguistics, 25*(1), 62-88. https://doi.org/10.1093/applin/25.1.62

Creswell, J. W. (2014). *Qualitative, quantitative and mixed methods approaches.* Sage.

Cruse, D. A. (1986). *Lexical semantics.* Cambridge University Press.

Davies, M. (2008-). *The corpus of contemporary American English (COCA): One billion words, 1990-2019*. Available online at https://www.english-corpora.org/coca/.

Edmonds, P., & Hirst, G. (2002). Near synonyms and lexical choice. *Computational Linguistics, 28*(2), 105-144. https://doi.org/10.1162/089120102760173625

Flowerdew, L. (2010). Using corpora for writing instruction. In McCarthy, M. & O'Keeffe, A. (Eds.), *The Routledge handbook of corpus linguistics* (pp. 444-457). Routledge.

Gabrielatos, C. (2005). Corpora and language teaching: Just a fling or wedding bells? Teaching English as a Second Language-Electronic Journal, *8*(4), 1-35. Retrieved August 2023, from http://tesl-ej.org/ej32/a1.htm.

Geluso, J., & Yamaguchi, A. (2014). Discovering formulaic language through data-driven learning: Student attitudes and efficacy. *ReCALL, 26*(2), 225-242. https://doi.org/10.1017/S0958344014000044

Gilquin, G. (2022). Cognitive corpus linguistics and pedagogy: From rationale to applications. *Pedagogical Linguistics*, *3*(2), 109-142. https://doi.org/10.1075/pl.22014.gil

Gilquin, G., & Granger, S. (2022). Using data-driven learning in language teaching. In A. O'Keeffe, M. J. McCarthy (Eds.). *The Routledge handbook of corpus linguistics* (pp. 430-442). Routledge.

Johns, T. (1986). Micro-concord: A language learner's research tool. *System, 14*(2), 151-162. https://doi.org/10.1016/0346-251X(86)90004-7

Karras, J. N. (2016). The effects of data-driven learning upon vocabulary acquisition for secondary international school students in Vietnam. *ReCALL, 28*(2), 166-186. https://doi.org/10.1017/S0958344015000154

Lee, H., Warschauer, M., Lee, J. H. (2019). The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis, *Applied Linguistics, 40*(5), 721-753, https://doi.org/10.1093/applin/amy012

Lei, L., & Liu, D. (2018). The academic English collocation list: A corpus-driven study. *International Journal of Corpus Linguistics, 23*(2), 216-243. https://doi.org/10.1075/ijcl.16135.lei

Leńko-Szymańska, A. (2022). Training teachers and learners to use corpora. In R. R. Jablonkai, E. Csomay (Eds.), *The Routledge handbook of corpora and English language teaching and learning* (pp. 509-524). Routledge.

Lin, Y. Y., & Chung, S. F. (2021). A corpus-based study on two near-synonymous verbs in academic journals: propose and suggest. *English Teaching & Learning, 45,* 189-216. https://doi.org/10.1007/s42321-020-00072-0

Liu, D., & Jiang, P. (2009). Using a corpus-based lexicogrammatical approach to grammar instruction in EFL and ESL contexts. *The Modern Language Journal, 93*(1), 61-78. https://doi.org/10.1111/j.1540-4781.2009.00828.x

Liu, D., & Zhong, S. (2016). L2 vs. L1 use of synonymy: an empirical study of synonym use/acquisition. *Applied Linguistics, 37*(2), 239-261. https://doi.org/10.1093/applin/amu022

Liu, D. (2018). A corpus study of Chinese EFL learners' use of circumstance, demand, and significant: An in-depth analysis of L2 vocabulary use and its implications. *Journal of Second Language Studies, 1*(2), 309-332. https://doi.org/10.1075/jsls.00006.liu

Mizumoto, A., & Chujo, K. (2015). A meta-analysis of data-driven learning approach in the Japanese EFL classroom. *English Corpus Studies*, *22*, 1-18.

O'Keeffe, A. (2023). A theoretical rationale for the importance of patterning in language acquisition and the implications for data-driven learning. *Nordic Journal of English Studies*, *22*(1), 16-41. https://doi.org/10.35360/njes.793

Pérez-Paredes, P. (2019). A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011-2015. *Computer Assisted Language Learning, 35*(1-2), 36-61. https://doi.org/10.1080/09588221.2019.1667832

Pérez-Paredes, P., Guillamón, C. O., Van de Vyver, J., Meurice, A., Jiménez, P. A., Conole, G., & Hernández, P. S. (2019). Mobile data-driven language learning: Affordances and learners' perception. *System, 84,* 145-159. https://doi.org/10.1016/j.system.2019.06.009

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing, 18*(1), 55-88.

Selivan, L. (2023). Corpus linguistics and vocabulary teaching. In K. Harrington and P. Ronan (Eds.), *Demystifying Corpus Linguistics for English Language Teaching* (pp. 139-161). Springer International Publishing.

Sinha, T. S. (2021). EFL learners' perception of and attitude to corpus as a vocabulary learning tool. *The Reading Matrix: An International Online Journal, 21*(2), 106-119.

Song, Q. (2021). Effectiveness of corpus in distinguishing two near-synonymous verbs: "Damage" and "destroy". *English Language Teaching, 14*(7), 8-20. https://doi.org/10.5539/elt.v14n7p8

Szudarski, P. (2019). Effects of data-driven learning on enhancing the phraseological knowledge of secondary school learners of L2 English. In P. Crosthwaite (Ed.), *Data-driven learning for the next generation* (pp. 133-149). Routledge.

Szudarski, P. (2022). Corpora and teaching vocabulary and phraseology. In R. R. Jablonkai, E. Csomay (Eds.), *The Routledge handbook of corpora and English language teaching and learning* (pp. 41-55). Routledge.

Yang, J., & Mei, F. (2024). Promoting critical reading instruction in higher education: A three-step training scheme facilitated by using corpus technology. *Journal of China Computer-Assisted Language Learning*, *4*(1), 15-142. https://doi.org/10.1515/jccall-2023-0029

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

Xiao, R., & McEnery, T. (2006). Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. *Applied Linguistics, 27*(1), 103-129. https://doi.org/10.1093/applin/ami045

Zareva, A. (2017). Incorporating corpus literacy skills into TESOL teacher training. *ELT Journal, 71*(1), 69-79. https://doi.org/10.1093/elt/ccw045

Authors

Sibel Söğüt, Ph.D., is an Assistant Professor at the English Language Teaching Department at Sinop University, Turkiye. She teaches undergraduate courses in the English language teacher training program. Her research interests are pre-service English language teacher training, critical pedagogy, second language writing, and data-driven learning. She has published in Computers and Education, Sustainable Development, TESL-EJ, and the Iranian Journal of Language Teaching Research.

ORCID ID: https://orcid.org/0000-0002-3395-7445.