

The Development of an Error-tagged Learner Corpus: TELC (Turkish-English Learner Corpus) and its Web-interface

Hakan Cangır¹, Kutay Uzun², Taner Can³
Enis Oğuz⁴, Ömer Faruk Kaya⁵

ORCID: ¹0000-0003-2589-2466, ²0000-0002-8434-0832,
³0000-0001-8869-4817, ⁴0000-0001-5819-4926
⁵0000-0001-7329-5557

¹Ankara University, School of Foreign Languages, Ankara

²Trakya University, Faculty of Education, Edirne

³TED University, Department of English Language and Literature, Ankara

⁴METU, Department of Basic English, Ankara

⁵Bursa Technical University, Project Support Office, Bursa

¹hcangir@ankara.edu.tr, ²kutayuzun@trakya.edu.tr, ³taner.can@tedu.edu.tr,
⁴enisoguz@metu.edu.tr, ⁵omer.kaya@btu.edu.tr

(Received 24 May 2024; Accepted 3 October 2024)

ABSTRACT: Though rather rare and not favoured by corpus linguists due to computationally hard-to-handle problems, learner corpora consisting of spoken and written texts by students from different L1 backgrounds can benefit both researchers in the field of second language acquisition and language teachers. Growing from this need and considering corpora's potential importance for the language teachers and learners in the Turkish context, our L2 English learner corpus is yet another humble attempt to build an error-tagged learner corpus particularly scrutinizing lexical errors, which play a key role in the language production of second language learners. Building on Hemchua and Schmitt's lexical error taxonomy and developed following the strict methodological considerations in the literature (e.g., error naming and fixing through several rounds of tagging), the corpus consists of 369 written texts by 231 university students (with 104,864 words, 3000+ tagged and fixed errors). The corpus database is provided with a user-friendly web-interface, which consists of statistical output, modules highlighting lexical errors and correct versions, different search options including error types, and an error-tagging add-in for further development. In addition to being a resourceful website trying to guide

language practitioners and second language learners, it can be considered a platform with a capacity to be developed further by applied linguists conducting studies in this line of research. Finally, thanks to its easy-to-use interface and versatile features, it has potential to become a reference learner corpus for English as a foreign/second language with the contribution of other universities in Türkiye.

Keywords: learner corpus, error-tagging, lexical errors, second language acquisition

Hata Etiketli Öğrenen Derlemi Geliştirilmesi: TELC (Türkçe-İngilizce Öğrenen Derlemi) ve Web-Arayüzü

ÖZ: Oldukça nadir olmasına ve derlem dilbilimciler tarafından geliřtirmedeki zorlukları nedeniyle tercih edilmemesine rağmen, farklı D1 gemişlerine sahip öğrencilerin sözlü ve yazılı metinlerinden oluşan öğrenen derlemleri, hem ikinci dil edinimi alanındaki arařtırmacılara hem de dil öğretmenlerine fayda sağlayabilir. Bu ihtiyaçtan yola çıkarak ve derlemlerin Türkiye bağlamında dil öğretmenleri ve öğrenenler için potansiyel önemini göz önünde bulundurarak, D2 İngilizce öğrenen derlemimiz, özellikle ikinci dil öğrenenlerin dil üretiminde kilit rol oynayan sözcük hatalarını inceleyen, hata etiketli bir öğrenen derlemi oluřtırmaya yönelik bir girişimdir. Hemchua ve Schmitt'in sözcüksel hata taksonomisine dayanan ve alanyazındaki katı metodolojik hususlar (örneğin, hata adlandırma ve birkaç tur etiketleme yoluyla düzeltme) izlenerek geliştirilen derlem, 231 üniversite öğrencisinin 369 yazılı metninden (104.864 sözcük, 3000'den fazla etiketlenmiş ve düzeltilmiş hatadan) oluşmaktadır. Kullanıcı dostu arayüze sahip derlem veri tabanı, kullanıcıların istatistiksel çıktılara ulaşmasına ve sözcüksel hataları ve doğru versiyonlarını görüntüleyebilmesine ve derlem içinde farklı hata türlerini aramasına imkân sağlar. Ayrıca, arayüzde veri tabanının gelişimine olanak sağlayan hata etiketleme eklentisi mevcuttur. TELC, dil öğretmenlere ve ikinci dil öğrenenlere rehber kaynak niteliğinde bir internet sitesi olmasının yanı sıra, bu alanda çalışmalar yürüten uygulamalı dilbilimciler tarafından geliştirilebilecek bir dijital platform olarak da değerlendirilebilir. Son olarak, kullanımı kolay arayüzü ve çok yönlü özellikleri sayesinde, Türkiye'deki diğer üniversitelerin de katkısıyla yabancı/ikinci dil olarak İngilizce öğretimi / öğrenimi için referans bir öğrenen derlemi olma potansiyeline sahiptir.

Anahtar Sözcükler: öğrenen derlemi, hata işaretleme, sözcük hataları, ikinci dil edinimi

1 Introduction: Learner Corpus Research

As a method of analyzing texts stored in electronic forms, corpus emerged as a distinct field in the 1960s with the emergence of the first modern corpus, *Brown Corpus of American English* (Francis & Kučera, 1964; Kučera & Francis, 1967).

Since then, the field of corpus linguistics has been contributing to language research and pedagogy with a richness of data and tools that are continually growing and improving. The early studies were predominantly concerned with L1 English varieties, which led to influential pedagogical reference books such as Collins *COBUILD English Grammar* (Sinclair, 1990). Although authentic L1 English data can give information on what is typical in English, there was also a need for a database that shed light on the characteristics and the needs of English language learners. The era of learner corpus research (LCR) began with the *International Corpus of Learner English* (ICLE; Granger, 1993), which started as a project to collect and study the writings of advanced learners of English as a foreign language. The systematically collected learner language data came to be known as learner corpus, which is “electronic collections of natural or near-natural data produced by foreign or second language (L2) learners and assembled according to explicit design criteria” (Granger et al., 2015, p. 1). The International Corpus of Learner English (ICLE) has evolved over time. It started small, with the first version containing 2.5 million words from students with 11 different native languages. This was all stored on a CD-ROM. The second version expanded to 4 million words, covering 16 native languages, and importantly, added grammatical tagging to the data. Now, the latest version (ICLEv3) is available online and boasts a massive 5.5 million words from students representing 25 native languages. (Granger et al., 2020). The dynamic progression of the *ICLE* project over the years mirrors the transformation of learner corpus research from a novice field, an “offshoot” of corpus linguistics, to a well-established discipline on its own. As reflected in *The Cambridge Handbook of Learner Corpus Research*, *The Journal of Learner Corpus Research* and many other LCR studies, the field has matured in terms of the availability of corpora, the quality of learner corpus studies, and the growing rapprochement of SLA and LCR.

Traditionally SLA researchers relied on smaller samples of data because the manual collection and analysis of large-scale samples were labor-intensive and time-consuming until recently (Granger, 2002). One of the main contributions of LCR is the volume and variety of learner data it has made available to the research community pursuing theoretical or applied research on L2 acquisition (Gilquin & Granger, 2015). Though early learner corpora (or L2 corpora) largely focused on written production (particularly argumentative and narrative texts) of learners collected in cross-sectional designs (see Gilquin, 2015 for an overview), online repositories now provide access to more than 200 learner corpora (see *Learner Corpora Around the World by UClouvain*¹) that come in different sizes and designs including longitudinal and cross-sectional, and modalities (i.e.,

¹ For the full list, see: www.uclouvain.be/en-cecl-lcworld.html (last accessed on 4 August 2024)

spoken and written). A recent example, *The Varieties of English for Specific Purposes Database* (VESPA; Paquot et al., 2022), is a multi-discipline and multi-register large-scale (over 2 million words) learner corpora, which is compiled of academic writings from university students with different L1s. The corpus includes samples of learner language from different disciplines (e.g., literature and business communication), academic levels, and registers (e.g., academic proposals), providing a wider perspective into characteristics of academic writing by non-native students. Publishing houses and testing organizations also created learner corpora, which are marginally larger than those that were compiled by academic circles. *Cambridge Learner Corpus* by Cambridge University Press, for example, is one of the largest learner corpora containing over 40 million words of exam scripts collected from university students. After compilation, learner corpora are often accompanied by added layers of information (e.g., error labels and syntactic annotation) to the raw data to allow researchers to investigate any language feature of interest and the context in which they appear. For instance, half of the *Cambridge Learner Corpus* is error-coded and all the texts in the corpus are also fully POS-tagged, which enables investigations into morphological acquisition (as in Murakami & Alexopoulou, 2016). Indeed, most of the encoding of the text is done by semi-automatic or fully automatic software tools. Lack of tools support became a thing of the past after the development of powerful computers and precise corpus tools. To illustrate, the webpage *Tools for Corpus Linguistics* (Berberich & Kleiber, 2023) features a list of 280 tools currently used in corpus linguistics research. These tools not only help the users to compile or annotate corpora, but also analyze them. Antconc (Anthony, 2023) is one of the most popular programs for visualizing frequency information and reviewing concordance lists, which shows the queried keywords in context. There are also tools for more complex analyses such as TAALES (Kyle et al., 2018) and TAASC (Kyle, 2016) for profiling lexical and syntactic sophistication by calculating dozens of related measures and indices.

In line with the advancements in corpus design, corpus annotation, and automated data extraction, the LCR field has also matured in how it analyses learner corpora. Due to the initial excitement over accessing previously inaccessible frequency information, many studies primarily focused on comparisons and productions of frequency lists for specific language features (e.g., verbs) and they made little use of tools other than concordancing software. According to Meunier (2020), most studies at the time could not go beyond the analysis of overused and/or underused linguistic items, generating lists of top n words in a corpus, and the documentation of the most frequently used linguistics phenomena. In Meunier's words, this led to "descriptive fever", a focus of interest or emphasis on the description of learner language rather than explaining what affects its development. LCR thus contributed more to empirical learner

language description than understanding the L2 knowledge that underlies language learning (Myles, 2005). The findings of these descriptive studies, while meaningful, have led to criticism of LCR for being merely descriptive and lacking critical analysis. However, as Myles argued, good descriptions of learner language can lay a solid foundation for understanding factors contributing to its development. Over the years, there has been a shift towards more exploratory and theory-driven investigations taking a wider range of variables into account. The potential of big empirical learner data to validate or challenge SLA theories has already been demonstrated by several studies (e.g., Biber et al., 2011; Murakami & Alexopoulou, 2016). One of the fundamental goals of SLA research is to establish if and when a certain structure is acquired (Bley-Vroman, 1989). To address this, Murakami and Alexopoulou (2016) analysed the *Cambridge Learner Corpus* to investigate the L2 acquisition order of six English grammatical morphemes (articles, past tense -ed, plural -s, possessive 's, progressive -ing, and third person -s) by learners from seven L1 groups across five proficiency levels. They found L1 influence on the absolute accuracy of morphemes and their acquisition order, therefore demonstrating that the accuracy order of L2 English grammatical morphemes is not universal but varies across learners with different L1 backgrounds. The study established a clear L1 influence on the absolute accuracy of morphemes and their acquisition order, therefore challenging the theories of the universal order of acquisition of L2 morphemes and informing the markedness theories of SLA.

Overall, learner corpus research has seen significant improvements in corpus documentation, design, and analysis, bridging the gap between research on SLA and LCR (Myles, 2021). LCR is inherently an interdisciplinary approach lying at the crossroads of corpus linguistics and SLA. As pointed out by Granger (2021), however, there are many other theoretical, methodological, and applied issues and challenges yet to be overcome through a cross-perspective approach and greater collaboration.

Having explored the rich data available in learner corpora, we now turn our attention to the practical application of this resource. The insights gleaned from learner corpora hold significant potential for informing and improving language teaching methodologies. Section 2 will examine how educators can leverage learner corpora to design effective instructional materials, address specific student errors, and ultimately enhance the overall language learning experience.

2 Learner Corpora for Language Teaching

Corpora have been a practical way to introduce authentic language use in language teaching, which is seen as the ultimate goal in theories such as data-driven learning (Pérez-Paredes, 2022). Despite their great potential, however, they have been mostly ignored in the creation of language education materials,

and language textbooks are generally based on the language assumed to be used in real life (O'Keeffe et al., 2007, p. 21). The biggest advantage of using corpora lies in its authentic nature, namely, allowing an objective investigation of how a certain language is used in real-life situations. This allows a thorough examination of word frequency, collocations, lexical variation, lexis in grammar, and authenticity (Hunston, 2002, p. 96). These generally acclaimed advantages have also been criticized as laying too much importance on frequency or authenticity in de-contextualized contexts and argued to undermine the importance of more specific but low-frequency words, such as those with high cultural values (Hunston, 2002, p. 194-195). However, measuring word frequency has been regarded as only one of the important aspects of corpora (Xiao, 2009); no language educator would argue in favour of presenting it as the sole criterion for achieving authentic language use. Therefore, the use of corpora goes beyond checking the frequency of words and phrases.

Despite the necessity of some technical knowledge, the use of corpora in language classrooms largely depends on pedagogical knowledge and expertise, and failing to accommodate such requirements would result in no positive outcomes from a method that could otherwise bring substantial benefits (Lee, 2011). Although L1 corpora can be a useful tool in studying the target language in an authentic way and in checking the usability of certain phrases and structures within the rules of target grammar, examining how L2 users produce the language is also essential in shaping second language education (Gilquin, 2023).

Learner corpora include “attempts” to use the target language as well as acceptable language usage, thus differing from L1 corpora, which consists of natural and mostly error-free language data. In general, how learners attempt to use language and make mistakes is based on teacher or language specialist intuitions. Learner corpora, on the other hand, present authentic error patterns of learners in a more systematic way and thus allows language educators to take necessary precautions and shape their learning according to empirical data (Thewissen, 2015). Comparing the usage of an L1 corpus to the usage of an L2 corpus can be a useful method in language classrooms (Xu, 2016), as it can help learners discover the mistakes made by L2 users like themselves and promote learning autonomy (Kaya et al., 2022; Nesselhauf, 2004 p. 140).

As mentioned earlier, learner corpora can also show considerable variance within themselves; therefore, it is not possible to consider all learner corpora as having similar characteristics, and the individual characteristics of a corpus (L1 or L2) should be carefully examined before suggesting pedagogical implications (Gablasova et al., 2017). The most essential learner characteristics include L2 proficiency, L2 exposure, context, and L1 background, as such differences can lead to different expertise in using certain structures and different error patterns. Keeping such variables similar across participants would allow researchers and

language educators to scrutinize the error patterns of learners at different proficiency levels accurately.

While some individual characters can be constant in a learner corpus, others may be allowed to vary in a systematic way. L1 background is perhaps the most common constant. The use of corpora has the potential to reveal much-neglected aspects of L1 influence and shed light on the sources of learner errors in L2 (Paquot & Granger, 2012), and learner corpora are great sources for investigating L1 influence on L2 language production (Granger, 2003). For instance, this is crucial for correcting collocation errors, as they mainly seem to stem from L1 influence (Nesselhauf, 2005; Laufer & Waldman, 2011). In addition, L1 can also lead to overuse or underuse of certain L2 phrases (e.g., Liao & Fukuya, 2004), and learner corpora are great tools to pin down such patterns (Paquot & Granger, 2012). However, if the L1 backgrounds of the participants vary, the investigation of L1 influence would only result in a misleading and blurry picture. Another common constant in such corpora is the context in which L2 was acquired, which can affect L2 acquisition (see Ellis & Laporte, 2014). L2 users may acquire their L2 in a natural context by getting exposed to rich authentic L2 input in terms of both quality and quantity. Other L2 users acquire their L2 mainly in a classroom setting, limiting their L2 interactions to mostly inauthentic language usage. Such differences have a strong potential to influence L2 acquisition speed and quality, error patterns, and the authenticity of the language used; therefore, focusing on a single context is likely to reveal a much clearer picture.

As mentioned above, some individual characteristics of a learner corpus can be allowed to show variety in a systematic way. L2 proficiency, for example, may vary across the sample pool of a corpus, but if this is not done in a systematic way in which each participant's L2 proficiency is determined and indicated in the corpus, the influence of L2 proficiency might disallow an accurate investigation of L1 influence and L2 acquisition context (Granger, 2015). When such an index is clearly reflected in the corpus, researchers can examine how constants (e.g., L1 influence) show different effects on error patterns across varying L2 proficiency levels. This knowledge is also helpful in language classrooms since L2 errors usually follow predictable patterns across different proficiency levels (Thewissen, 2013); teachers can expect certain errors from certain proficiency levels, and they can get ready for such errors even before observing them in student texts. Furthermore, students can examine the text written by others with similar L2 proficiency, recognize mistakes, and compare them with their own essays. The developmental map of L2 learners would also be helpful in identifying what is teachable at different L2 proficiency levels, as this should be taken in consideration while designing lesson materials (Ellis & Laporte, 2014).

The final consideration when using learner corpora in language education is text characteristics. Texts in a learner corpus can vary in terms of formality,

genre, length, and topic (Granger, 2002). The difference between formal and informal writing can be challenging for an L2 learner at times; L2 learners tend to use more informal language in their formal writing tasks (Lee et al., 2019). While a learner corpus focusing merely on formal texts can reveal common formality mistakes, if the corpus takes formality as a variable and presents both formal and informal texts written by the same sample of L2 learners, researchers and language educators can also analyze which mistakes are transferred from informal L2 writing tasks and which are created in the process of writing formal L2 writing tasks.

In summary, learner corpora stand as a beneficial tool in language education due to their authentic representation of L2 use and errors in a systematic way. In order to take full advantage of these tools, participant and text characteristics of a learner corpus should be carefully examined, and the patterns observed should be evaluated within pedagogical frameworks before being implemented into language education.

Emerging from the stated importance of a learner corpus including texts of students from different L1 backgrounds (e.g., Turkish) and the growing need for learner corpora in the Turkish context, the present study sets out to build a balanced and representative learner corpus highlighting the lexical errors made by the university students in their argumentative essays. The developed corpus is released through a versatile web interface with unique features, such as error search, textual analysis (for frequency, grammar errors etc.) and text comparisons.

3 TELC Corpus Design

3.1 Error-Coding

3.1.1 Coders

Our error-coding team consisted of six members. Five of them were academicians in language science departments at different universities in Türkiye, and one of the members was an MA student in the English Language Teaching department. Three of the academicians hold PhD degrees in language teaching, literature, and linguistics. Two of them had their PhD education in language teaching and literature departments during the coding phase. All the members have experience in teaching English as a foreign language at university level (ranging from 2 to 18 years).

3.1.2 Error-taxonomy

The present investigation adopts Hemchua and Schmitt's taxonomy (2006) as the foundation (Table 1). This taxonomy is comprehensive, drawing from previous

classifications (e.g., Leech, 1981), and encompasses all the primary lexical errors that have been examined to date. To enhance practicality and ensure greater reliability during the error coding phase, we consider all the sub-headings during error detection. However, in the error tagging phase, we restrict our usage to the main headings of the taxonomy. Our decision is driven by practical considerations, anticipating that the final error-tagged corpus will primarily benefit language instructors and English learners. We aspire to present potential users with a more user-friendly interface and error types that are pedagogically convenient and suitable. Building upon earlier studies (e.g., Granger, 2003) and guided by our preliminary analysis of student essays, we expanded the main headings to incorporate major sources of lexical errors specific to the Turkish context. The modified and extended version of the taxonomy is displayed in Table 2. While the definition of an error might differ between prescriptive and descriptive approach, we adopted a natural-language-use approach and investigated whether specific patterns are acceptable and commonly used in native language production using L1 corpora and dictionaries, in addition to consulting native speakers.

Table 1. Hemchua and Schmitt's error taxonomy

FORMAL ERRORS (FE)	SEMANTIC ERRORS (SE)
FORMAL MISSELECTION (FM)	1 CONFUSION OF SENSE
1.1 suffix type	RELATIONS (SR)
1.2 prefix type	1.1 general term when a specific one is required
1.3 vowel-based type	1.2 overly specific term
1.4 consonant-based type	1.3 inappropriate co-hyponym
1.5 false friends	1.4 near synonyms
2 MISFORMATIONS (MI)	2 COLLOCATION ERRORS (CL)
2.1 borrowings	2.1 semantic word selection
2.2 coinage	2.2 statistically weighted preferences
2.3 calque	2.3 arbitrary combinations & irreversible binominals
	2.4 preposition partners
3 DISTORTIONS (DT)	3 CONNOTATION ERRORS (CE)
3.1 omissions	
3.2 overinclusion	
3.3 misselection	
3.4 misordering	
3.5 blending	
	4 STYLISTIC ERRORS (SY)
	4.1 verbosity
	4.2 underspecification

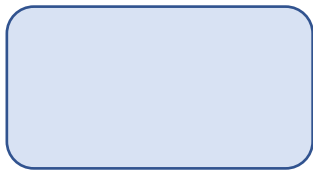
In Hemchua and Schmitt's (2006) error taxonomy, semantic lexical errors occur when a learner uses a word with an incorrect meaning, often due to the confusion between semantically related words (e.g., using "borrow" instead of "lend"), wrong choice of word combinations or unnatural chunks (e.g., "heavy coffee" instead of "strong coffee"), connotational confusions (e.g., the use of "effect" with a negative connotation) and stylistic errors (e.g., "authorised people" instead of "authorities").

On the other hand, formal lexical errors involve mistakes related to the form of the word due to its spelling (e.g., "congection" instead of "congestion"), morphological form (e.g., "economic" instead of "economical") or coinage (e.g., "solving" instead of "solution").

3.1.3 Adapting and calibrating the lexical error-coding scheme: Coding, naming and correcting

3.1.3.1 Semi-automated error coding procedure

We utilized the error-tagging add-in within Google Docs, as illustrated through a representative sketch in Figure 1. A custom code was developed for this add-in, facilitating the identification and selection of errors. The code generated error tags and their corresponding correct versions, enabling the seamless replacement of the erroneous words or phrases. To ensure compatibility with Sketch Engine (Kilgarriff et al., 2014), we adhered to the coding scheme provided by Sketch Engine. This choice allows for the automatic recognition of our annotations by Sketch Engine and the subsequent automatic construction of our error-tagged learner corpus on the corpus website (TELC).

Error Labelling Steps	Replacement																				
<ol style="list-style-type: none"> 1. Select erroneous text, 2. Choose error type, 3. Press <i>Generate</i> button, 4. Add correction in between corr. tags, 5. Press <i>Replace</i> button, 6. Error count updates automatically. 																					
Error Types																					
Formal Lexical Errors <ul style="list-style-type: none"> ▪ Formal misselection (affix type, vowel consonant based) ▪ Misformations (borrowing, coinage, calque) ▪ Distortions (spelling based) 	<div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid black; border-radius: 10px; padding: 5px; background-color: #e0e0e0;"><i>Generate</i></div> <div style="border: 1px solid black; border-radius: 10px; padding: 5px; background-color: #e0e0e0;"><i>Replace</i></div> </div>																				
Semantic Lexical Errors <ul style="list-style-type: none"> ▪ Confusion of sense relations (too general/specific, near synonyms) ▪ Collocation errors ▪ Collocation errors (L1 transfer) ▪ Collocation errors (due to preposition) ▪ Connotation errors ▪ Stylistic errors (verbosity, underspecification, formal/informal use) 	<table border="1"> <thead> <tr> <th colspan="2">Error Counts</th> </tr> </thead> <tbody> <tr><td><i>fm</i></td><td style="text-align: right;">0</td></tr> <tr><td><i>mi</i></td><td style="text-align: right;">0</td></tr> <tr><td><i>dt</i></td><td style="text-align: right;">0</td></tr> <tr><td><i>sr</i></td><td style="text-align: right;">0</td></tr> <tr><td><i>cl</i></td><td style="text-align: right;">0</td></tr> <tr><td><i>clt</i></td><td style="text-align: right;">0</td></tr> <tr><td><i>clp</i></td><td style="text-align: right;">0</td></tr> <tr><td><i>ce</i></td><td style="text-align: right;">0</td></tr> <tr><td><i>sy</i></td><td style="text-align: right;">0</td></tr> </tbody> </table>	Error Counts		<i>fm</i>	0	<i>mi</i>	0	<i>dt</i>	0	<i>sr</i>	0	<i>cl</i>	0	<i>clt</i>	0	<i>clp</i>	0	<i>ce</i>	0	<i>sy</i>	0
Error Counts																					
<i>fm</i>	0																				
<i>mi</i>	0																				
<i>dt</i>	0																				
<i>sr</i>	0																				
<i>cl</i>	0																				
<i>clt</i>	0																				
<i>clp</i>	0																				
<i>ce</i>	0																				
<i>sy</i>	0																				

*Source for XML Coding (Sketch Engine): https://www.Sketch_Engine.eu/documentation/setting-up-learner-corpus/#toggle-id-1

Figure 1. A sketch of Google document ad-in for error annotation

3.1.3.2 Error naming phase

Our lexical error taxonomy choice (Hemchua & Schmitt, 2006) aimed to align with our research requirements and had the potential to encompass context-specific errors. While considering sub-headings in the labeling of detected lexical errors, we opted for the main headings (and disregard the sub-headings) of Hemchua and Schmitt's taxonomy to ensure reliability and consistency in error naming. This decision aimed to create a more accessible error-tagged corpus for prospective users such as teachers, students, and researchers, following Granger's (2002, 2003) recommendation for taxonomies to be comprehensive yet manageable. Following initial analysis and calibration meetings, where we worked with sample student papers, we incorporated context-specific errors like L1-based collocation errors and prepositional (collocation) errors into our

taxonomy. This addition was motivated by the prevalence of transfer errors, particularly in word combinations and the distinct syntactic structures of Turkish and English, such as the use of postpositions in Turkish instead of prepositions. To prepare annotators, we invested time in training sessions. During calibration meetings, we engaged with sample student papers, initially locating errors and subsequently naming them according to our error taxonomy.

Table 2. Simplified and enhanced version of error taxonomy

FORMAL ERRORS (FE)	SEMANTIC ERRORS (SE)
<ul style="list-style-type: none"> ▪ FORMAL MISSELECTION (FM) 	<ul style="list-style-type: none"> ▪ CONFUSION OF SENSE RELATIONS (SR)
<ul style="list-style-type: none"> ▪ MISFORMATIONS (MI) 	<ul style="list-style-type: none"> ▪ COLLOCATION ERRORS (CL) ▪ COLLOCATION ERRORS DUE TO L1 (CLT) ▪ COLLOCATION ERRORS DUE TO PREP. (CLP)
<ul style="list-style-type: none"> ▪ DISTORTIONS (DT) 	<ul style="list-style-type: none"> ▪ CONNOTATION ERRORS (CE) ▪ STYLISTIC ERRORS (SY)

Utilizing the final version of the error taxonomy outlined in Table 2, during the initial round of error coding, we employed pairs of annotators (a total of 6 annotators), with each pair independently coding the same texts. The Google document's automated feedback feature was considered for an initial analysis of potential errors. Notably, we found that Google's suggestions on grammar often pointed to lexical issues, such as collocational errors and spelling mistakes. Focusing on lexical errors, we disregarded grammar-related issues. When identifying potential lexical errors, we verified their validity using various resources, including reference corpora (e.g., COCA), online dictionaries (e.g., Longman Dictionary), corpus tools (e.g., JusttheWord), and specific Google searches (e.g., "traffic increase" the guardian). Potential L1 transfer errors were determined through backtranslation. If an error was tagged as a possible L1 transfer collocational error, we back-translated the collocation and verified its accuracy through the Turkish National Corpus (TNC - <https://www.tnc.org.tr>) and the Turkish Language Council's official online dictionary (<https://sozluk.gov.tr>).

After the first and second pairs completed their tagging, the final pair provided comments on the already tagged errors or identified additional potential errors. Group meetings were held every other week to go over the errors tagged identically by the pairs and discuss the disagreements. To enhance the reliability of error tags, we prioritized errors on which the majority of the team (at least 4 out of 6 annotators) reached a consensus regarding error type, while disregarding

potential errors with fewer than three agreements. This initial coding round served as a preliminary step or post-calibration phase to formulate the error-coding guide for subsequent use, establishing fundamental principles and ensuring annotators shared a common understanding of error locations and naming conventions. Based on the discussions and agreed-upon error examples, we designed an error guide to aid in the further tagging process and assist potential users of the learner corpus.

In the second round of error coding for the same texts, we adopted a slightly different approach. Two groups of researchers (3 members in each group) tagged a specific number of texts weekly. The focus was on reviewing the already tagged texts to refine lexical errors if necessary and identify any oversights. For challenging errors, the team sought assistance from a native speaker. This second coding attempt allowed us to fine-tune errors, align more closely with the error-coding guide and principles established in the first round, and address difficulties in naming certain error types. The final error tags were agreed on by all the members of the research team.

3.1.3.3 Error-correction phase

As stated earlier, we decided intuitively about the correction of the misuse and then we consulted native speaker corpora and dictionaries as well as a native speaker when necessary to consolidate expert intuitions. Although some sentences were syntactically problematic and impossible to fix without rewriting, we corrected some micro level lexical errors to make the sentences more meaningful. In other words, a sentence with some lexical error fixes could still be syntactically defective, sound unnatural, and may not be fully compositional. When an error correction required a change of more than 4 content words, we decided not to tag and fix that error as replacing that many words meant writing the sentences from scratch. We thought it was beyond our investigation and research scope. In other words, our lexical error tags, and corrections consisted of lexical items no more than four words at a time. We only violated the 4-word limit for stylistic errors in very few instances because particularly the verbosity errors (i.e., use of excessive words to express a simple concept) required such corrections. When there was more than one possibility for an error correction, we stuck to the correction requiring fewer word changes. While correcting errors, we inserted some grammatical words such as articles into our corrections. (e.g., make dishes - do THE dishes). Sample errors and corrections can be seen in Appendix A.

Error tags can be analysed in more detail using the simple learner corpus interface at www.telcorpus.org. Some sample errors and corrections are

presented in Table 3. More error examples can be reached at the error guide designed throughout the error-tagging phase.²

Table 3. Lexical error samples

Formal Lexical Errors	
Error type	Example
<ul style="list-style-type: none"> ▪ Formal misselection [fm] (due to suffix, prefix, vowel and consonant based errors) ▪ Misformations [mi] (borrowing, coinage and calque) ▪ Distortions [dt] (letter omissions, overinclusion, misselection, misordering, and blending) 	<ul style="list-style-type: none"> ▪ fastly – correction: fast (suffix) ▪ solving – correction: solution (coinage) ▪ succes – correction: success (missing letter)
Semantic Lexical Errors	
Error type	Example
<ul style="list-style-type: none"> ▪ Confusion of sense relations [sr] (too general, specific and near synonym) ▪ Collocations errors [cl] (statistically weighted preferences and arbitrary combinations) ▪ Collocational errors due to L1 [clt] (L1 transfer related) ▪ Collocational errors due to prepositions [clp] (wrong, under or overuse of prepositions) ▪ Connotation errors [ce] ▪ Stylistic errors [sy] (verbosity and underspecification) 	<ul style="list-style-type: none"> ▪ promote – correction: encourage (near synonym) ▪ intensive traffic – correction: heavy traffic (statistically weighted preferences) ▪ do a class – correction: attend class (L1 Turkish transfer: <i>ders yapmak</i>) ▪ attend to – correction: attend (overuse) ▪ popular – correction: serious (used with negative connotation) ▪ authorised people – correction: authorities (verbosity)

3.2 The Corpus

Once the parameters for error-naming and correction were set, the essays were initially semi-automatically tagged using the Sketch Engine’s error coding feature. The numerical output of the corpus can be seen in Table 4. The corpus is comprised of 369 texts written by 231 students who were majoring in English Language Teaching; English Language and Literature departments (aged between 19 and 35). Additionally, tertiary level students with at least an intermediate level of English proficiency took part in the study. The participants

² See “[Error Guide](#)” for more examples.

(63% female and 37% male) wrote about four different argumentative essay questions (see Appendix B for the selected questions), randomly selected from a pool of essay questions at a state university. The students submitted their timed assignments through an online platform where they did not have access to dictionaries or other online aid. Following the taxonomy presented in Table 2, the coders detected 3014 lexical errors, the dispersion of which is given in Table 5. Once the corpus is POS (part of speech) and error-tagged using the error-tagging scheme compatible with the Sketch Engine platform, it is possible to search for specific errors (see Figure 2). Using this feature, one can get the related concordance lines and further analyse the numerical output (see Figure 3) with raw and normalised values (e.g., hits per million words, lexical distribution etc.).

Table 4. Numerical output of the corpus

Corpus Feature	Numerical Output
Tokens	115,754
Words	104,864
Sentences	6,025
Documents	369

Table 5. Dispersion of lexical errors across error types

Error type	Frequency	Percentage
Confusion of sense relations (sr)	777	25,8%
Stylistic errors (sy)	568	18,8%
Collocational errors (cl)	466	15,5%
Collocational errors due to prepositions (clp)	464	15,4%
Formal misselection (fm)	225	7,5%
Distortions (dt)	207	6,9%
Collocational errors due to L1 transfer (clt)	157	5,2%
Misformations (mi)	123	4,1%
Connotational errors (ce)	27	0,9%

This versatile interface helps researchers visualize the corpus content, expand its content with meta-tagging features and provides a perfect platform to develop the corpus to be used for research and teaching purposes. However, as the Sketch Engine website provides a paid service for the potential users, we decided to develop a free and user-friendly learner corpus website with various search options and integrated text analysis tools embedded in the interface, particularly for the needs of language instructors.

The screenshot shows the 'ERROR ANALYSIS' tab of the Sketch Engine interface. It features a navigation menu with 'BASIC', 'ADVANCED', 'ABOUT', and 'ERROR ANALYSIS'. The main area includes a search bar with 'Error code' and a dropdown menu set to '== all =='. Below this are fields for 'Incorrect word(s)' and 'Correct word(s)'. A legend on the left allows users to select highlighting options: 'Highlight query result' (selected), 'Highlight error words', 'Highlight corrected words', and 'Highlight both'. A 'Text types' dropdown is also present. A red 'SEARCH' button is located at the bottom right. On the far right, a vertical list of error codes is visible, including 'ce', 'cl', 'clp', 'clt', 'dt', 'fm', 'mi', 'sr', and 'sy'.

Figure 2. Sketch Engine query interface

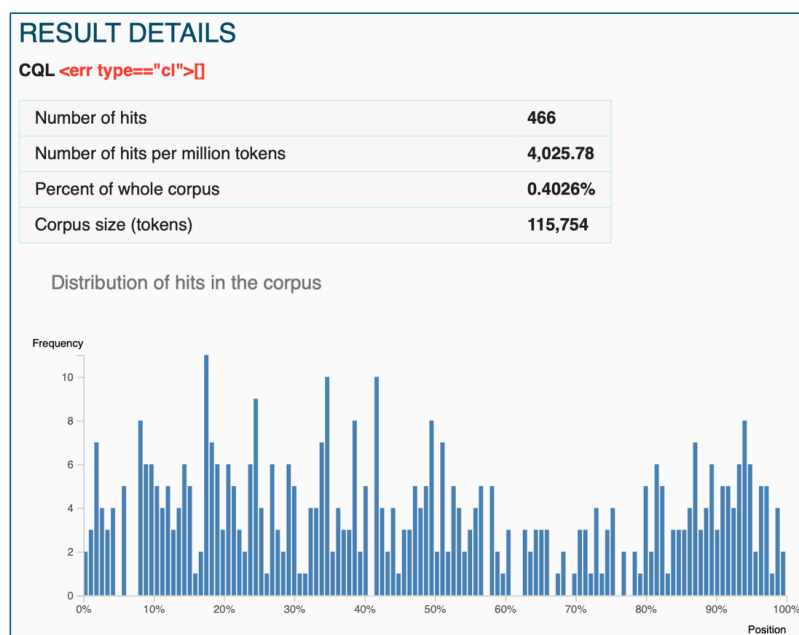


Figure 3. Sample CQL search on Sketch Engine

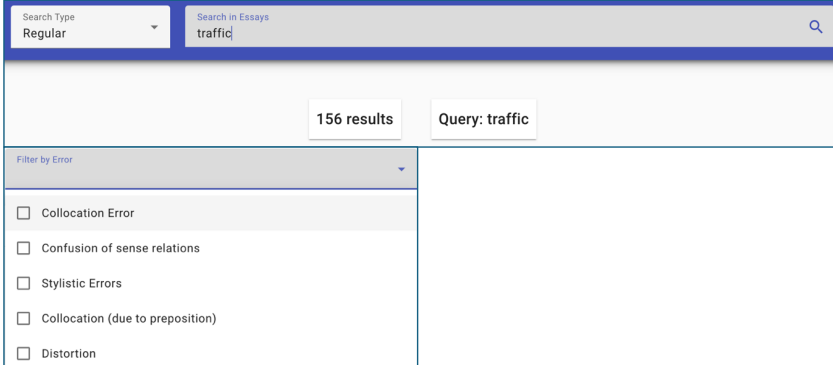
4 TELC Corpus Website

On the TELC corpus website, tagged lexical errors in texts can be analyzed, statistical findings can be examined, error counts, average performance scores given by human raters and scores generated by the score prediction model (with

a prediction accuracy of at least 80% - see Cangır et al., 2025 for further details) can be accessed. The first release of the corpus interface, which is open to further development, enables the addition of new texts with the error marking module. The corpus can be used by teachers and students for foreign language teaching. In the future, foreign language teaching materials can be created using the corpus and the corpus website can be further enriched in this way. This corpus website, which is the first of its kind in the field, has the potential to become a reference learner corpus in Turkey with the participation of other schools and institutions. The raw version of the corpus can be accessed at <https://sites.google.com/view/automated-grading/main-page>.

4.1 Features of TELC

Using the simple search option, users can search the token frequency for any word, word combinations and particular error tags. As Figure 4 shows, it is possible to filter the search by the type of error, allowing users to narrow down their focus on specific error types. Additionally, as is seen in Figure 5, the system provides concordance lines highlighting the target word and its surrounding context, which could help users detect some collocational patterns or observe some structural properties of the search terms (e.g., preposition that follow).



The screenshot displays the TELC search interface. At the top, there is a search bar with a dropdown menu for 'Search Type' set to 'Regular' and a search input field containing 'traffic'. Below the search bar, the results are summarized as '156 results' and 'Query: traffic'. A 'Filter by Error' dropdown menu is open, showing a list of error types with checkboxes: 'Collocation Error', 'Confusion of sense relations', 'Stylistic Errors', 'Collocation (due to preposition)', and 'Distortion'. All checkboxes are currently unchecked.

Figure 4. Query sample

Before	Query	After
...are lot of cars around there and this causes huge	traffic jam	s throughout the streets in city, especially in th...
...and leave the city, generally in that period, the	traffic jam	s less likely to happen. That's why i think there...
...e public transportation and this will relieve the	traffic jam	in the city and also it will contribute more to t...
...e in terms of eliminating the problems related to	traffic jam	. Still, the fact that most of the business centre...
...ars also steal our time by making us stuck in the	traffic jam	. Instead of using cars, taking the buses would be...
.... An average driver spends almost 99 hours in the	traffic jam	which is more than 4 days. Also, most people cann...
...ering others, or go over the speed limit to avoid	traffic jam	. Overall, the car usage in the big cities takes o...
...blems in big cities so using many cars can cause	traffic jam	and people go places where they want late. They c...
... they also cause some problems for our world like	traffic jam	, air and noise pollution and traffic accidents. B...

Figure 5. Sample concordance lines

Building on these basic features, the corpus also provides a platform where you can investigate errors on single student texts. This feature of the learner corpus makes essay comparisons possible (Figure 6).

Incorrect
Both
Correct

↔ Compare Side by Side

↔ Compare with Another Essay

<> Show Raw Essay

PRT_5_T1

The world's population continues to increase. Unfortunately, people's problems started to increase with the population, and one of these problems is **traffia** in city centers. Some people believe that cars should be banned from city centers to reduce traffic problems, especially in big cities. I personally believe that it is not a good idea, and I have several reasons.

Firstly, people might get hurt because of this rule. For example, if a family member of yours gets hurt **in the** outside of city center, and you do not have access to your mobile phone to call an ambulance, you cannot enter the center. Thus, people may not go to a hospital in limited time, and they might get hurt.

Secondly, we should not ban cars for **good**, but we should limit the number of cars in the city centre. For instance, a policeman can **control** how many car a family has, and a policeman can allow only one car in the same city center at the same time. Moreover, thanks to this, governments have to increase the number of policemen, and they can inspect the traffic much more than before. As a result, governments can restrict the number of cars in the city center.

Thirdly, the driver's license is too **flexible**. These days driver's license is easy to **reach**, so everybody can get it if they are **up to** the age of eighteen. Driver license courses should teach drivers better than now, and these courses should be harder than at present time.

To sum up, banning cars from city centers is not the best way to deal with the traffic problem. I personally believe that we have better alternatives to deal with it.

PRT_5_T1

The world's population continues to increase. Unfortunately, people's problems started to increase with the population, and one of these problems is **traffic congestion** in city centers. Some people believe that cars should be banned from city centers to reduce traffic problems, especially in big cities. I personally believe that it is not a good idea, and I have several reasons.

Firstly, people might get hurt because of this rule. For example, if a family member of yours gets hurt **none** outside of city center, and you do not have access to your mobile phone to call an ambulance, you cannot enter the center. Thus, people may not go to a hospital in limited time, and they might get hurt.

Secondly, we should not ban cars for **good**, but we should limit the number of cars in the city centre. For instance, a policeman can **check** how many car a family has, and a policeman can allow only one car in the same city center at the same time. Moreover, thanks to this, governments have to increase the number of policemen, and they can inspect the traffic much more than before. As a result, governments can restrict the number of cars in the city center.

Thirdly, the driver's license is too **easy to obtain**. These days driver's license is easy to **obtain**, so everybody can get it if they are **over** the age of eighteen. Driver license courses should teach drivers better than now, and these courses should be harder than at present time.

To sum up, banning cars from city centers is not the best way to deal with the traffic problem. I personally believe that we have better alternatives to deal with it.

Figure 6. Text comparison window with target words highlighted

To be more precise, users can analyze essays side by side with lexical errors being highlighted. When users hover their cursors over the highlighted items, they can see the corrected versions. They also have the option to highlight certain lexical errors and see the raw versions of the texts, which could be transferred to the Sketch Engine interface for further manipulation.

Below the text comparison windows, users can see the numerical and visual output (Figure 7) showing details regarding the number of words in the essay, the number and types of lexical errors, rater scores together with their average value and the model scores based on the score prediction model developed in Cangır et al., 2025 (with a prediction accuracy of 80%).



Figure 7. Number of errors and essay score ratings

The platform allows users to code errors in raw texts using its error-coding scheme. The coded texts are automatically detected by the Sketch Engine platform, and the researchers can easily transfer texts to Sketch Engine for further analysis.

On top of these features, the web interface comes with an essay analysis feature (e.g., word-for-word count, most frequent words, word type distribution, grammatical errors, word frequency and diversity components, etc.) in L2 texts using ready-made Python language libraries (Natural Language Toolkit, www.nltk.org and language_tool_python <https://pypi.org/project/language-tool-python/>) and provides users with the opportunity to measure text quality (Figure

8). It has a simple drag and drop window where users can copy and paste their texts and click on *Analyse Essay* to explore some textual features. The output window provides token counts, distinct token and word counts, content word counts (with their visuals) and the number of possible grammatical errors. These numbers can tentatively guide language practitioners regarding the quality of a text and can be used for research purposes. Since the software required for the automatic generation of some linguistic features (as suggested by Cangir et al., 2025) is not yet open source, these features could not be included in the module. In the future, if this software becomes open source, these linguistic features can also be included in the current software, and it can turn into an automated grading and feedback system with full functionality. Nevertheless, the model we created in the project could predict essay scores with over 80% accuracy; therefore, its possible integration can provide valuable information regarding user essay quality. Even if this feature becomes available, however, language practitioners should be aware that the model was created by using data from a specific group of university students, and its prediction might be limited when it comes to other L2 populations.

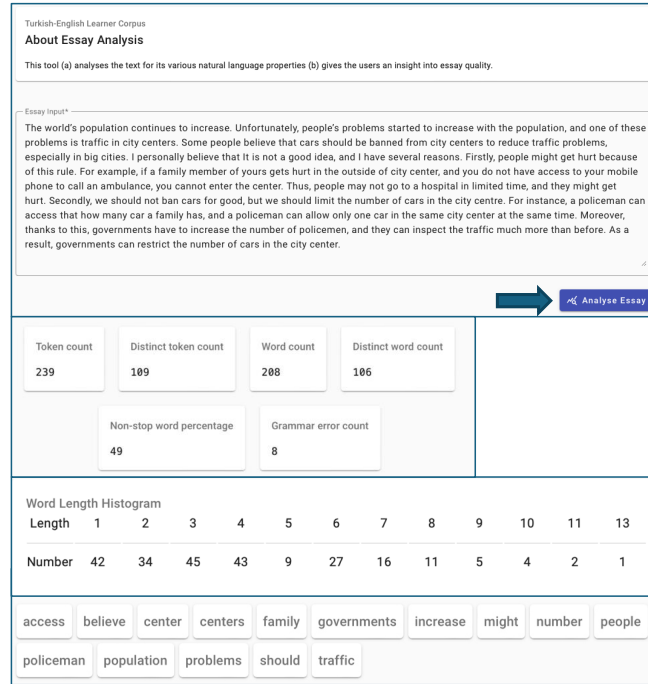


Figure 8. Essay analysis sample window

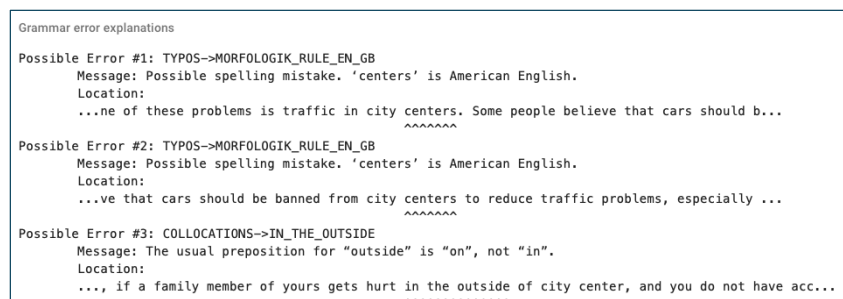


Figure 9. Sample window for grammar error explanations

Participant University: Ankara	Participant Gender: Female
Participant Age: 20	Participant Language Level: B1
Participant LexTale: 70	Participant LHQ Proficiency: 0.61
Participant LHQ Immersion: 0.28	Participant LHQ Dominance: 0.33

Figure 10. Sample screen for participant details

5 How can TELC be used for language teaching?

With its user-friendly interface, pedagogically rich features, and controlled development procedure, TELC can be used for language teaching and assessment purposes. First, it can directly be used by language instructors while teaching vocabulary in a classroom setting (Selivan, 2023). Common errors made by students at the tertiary level can be highlighted and explicitly discussed to raise awareness regarding the general lexical problems they are likely to face and to improve their metalinguistic skills (as also suggested by Paquot & Granger, 2012). There are studies in the literature (e.g., Schneider, 2023) emphasizing the benefit of using learner corpus (even without error tagging) to raise awareness about the common errors students make by analyzing sample student essays and concordances. In addition to single-word teaching, concordances extracted through the interface can be utilized to create language teaching materials in an academic writing classroom with a special emphasis on formulaic language use. The learners can analyze their own essays and get insight into possible grammar problems in their written production. The numerical output can guide the teachers in terms of the overall quality of the texts.

The platform can also be indirectly used by practitioners and material designers to create corpus-informed language teaching materials (Cortes, 2018) with special emphasis on the Turkish context. They can develop targeted materials, use learner corpora to create targeted exercises and materials that focus

on common problem areas (e.g., the *Common Mistakes at* series; Moore, 2005), and ensure that instructional content is more aligned with learners' needs (Ellis & Laporte, 2014). For instance, the (collocational) errors indicating potential L1 influence (Nesselhauf, 2003) can be given special attention to help learners discover the potential problematic areas in language use, and this emphasis can guide teachers into prioritizing certain lexical phenomena (Granger, 2003). Instructors can elicit common errors by analyzing sample concordances. This helps instructors tailor their teaching materials and strategies to address specific language challenges. As also suggested by Thewissen (2015), more teaching time could be needed particularly for incongruent word combinations (like collocations) in L1 and L2.

Learner corpora like TELC can benefit language specialists and applied linguistics indirectly (Gilquin, 2023). TELC can help them understand the interlanguage of the language users in the Turkish context, which refers to the transitional language stage learners go through. Analyzing learner corpora can help instructors understand the patterns and features of interlanguage, guiding them in providing appropriate support (Theweissen, 2013; Crosthwaite, 2024).

The platform can be used by the learners to foster self-reflection by comparing their essays with the essays provided on the website (as also suggested by Xu, 2016). By doing this, they can reflect on their own language use and identify patterns of errors or areas for improvement. Additionally, they compare their texts with proficient models, aiding in the development of a more native-like linguistic competence. Finally, using the learner corpus to discover the language patterns by themselves (Friginal, 2013), they take a more active role in their language learning, setting goals based on their individual needs and tracking their progress over time, which as a result encourages autonomous learning (Kaya et al., 2022).

In addition to enhancing language instruction, learner corpora like TELC can guide language assessment. The comparison of performance scores by human raters and a score prediction model (Cangır et al., 2025) can guide practitioners in terms of writing quality in L2 English. Though not available on the corpus website at the moment, an automated scoring module based on the underlying score prediction model and considering the linguistic features with strong predictive power can help teachers in their writing quality evaluations and help students write better essays.

In summary, incorporating TELC in language teaching is likely to facilitate a more data-driven and tailored approach for both instructors and students. It has the potential to enhance the understanding of learner needs, inform teaching strategies, and promote individualized language development.

6 Conclusion

In conclusion, as stated earlier, learner corpora can offer invaluable insights into the linguistic development of learners, making them a powerful tool for exploratory learning. By systematically analyzing learners' language usage patterns and their progression, educators can tailor their instruction to address specific learning gaps and reinforce areas of strength. Considering this potential and the unique features of TELC, the analysis of a lexical error-tagged learner corpus in the Turkish context holds significant implications for second language acquisition research and language pedagogy both in the Turkish context and globally. By identifying common lexical errors made by learners and having a deeper understanding of the language acquisition process, language learning facilitators can adjust instructional approaches to address specific linguistic challenges effectively. Additionally, the dynamic nature of learner corpora ensures that teaching strategies remain responsive to evolving learner needs, fostering a more effective and informed approach to language education. Furthermore, its potential to inform automated grading applications underscores its relevance in modern educational technology. However, it is essential to acknowledge that the current corpus may not be comprehensive enough (due to the limited number of texts and task types in its database) and calls for collaboration with other national universities to expand its scope. With concerted efforts, this corpus has the potential to evolve into a definitive reference for language learning and teaching in the Turkish university context, facilitating deeper insights into learners' linguistic development and pedagogical strategies. As technology continues to advance, we can anticipate even more sophisticated applications of learner corpora, further enhancing the effectiveness of language education.

Author Contributions: All authors contributed equally to the conceptualization, literature review, data collection, data analysis, and writing of this article.

Submission statement and verification: This study has not been previously published elsewhere. It is not under review in another journal. Publication of the study has been approved, either implicitly or explicitly, by all authors and the responsible authorities at the university/research center where the study was conducted.

Conflict of Interest Statement: The authors declare that there are no financial or academic conflicts of interest between themselves or with other institutions, organizations or individuals that may affect this study.

Data Use: Linguistic data reported in this study, unless a source is cited, come from authors' native speaker knowledge of Turkish, checked through the traditional method of informal consultation with other native speakers of Turkish.

Ethical Approval/Participant Consent: Ethical approval was obtained from Ethics Committee at Ankara University on March 3, 2012 (Decision no. 3/9).

Financial Support: This study is conducted as part of a research project supported by the Research Council of Türkiye, TÜBİTAK ARDEB (220K289).

References

- Anthony, L. (2023). AntConc (Version 4.2.4) [Computer Software]. Waseda University. Available from <https://www.laurenceanthony.net/software>
- Berberich, K., & Kleiber, I. (2023). *Tools for corpus linguistics*. <https://corpus-analysis.com/>
- Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *Tesol Quarterly*, 45(1), 5-35. <https://doi.org/10.5054/tq.2011.244483>
- Bley-Vroman, R. (1989). What is the logical problem of foreign language learning? In S. M. Gass & J. Schachter (Eds.), *Linguistic perspectives on second language acquisition* (pp. 41–68). Cambridge University Press. <https://doi.org/10.1017/CBO9781139524544.005>
- Cangır, H., Uzun, K., Can, T., Küllü, K., Oğuz, E., Kaya Ö. M. (2025). Linguistic features and L2 English writing quality: A multidimensional analysis. [Manuscript submitted for publication]. *AILA Review*.
- Cortes, V. (2018). Corpus tools for Writing Teachers. In *The TESOL Encyclopedia of English Language Teaching* (pp. 1–6). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118784235.eelt0553>
- Crosthwaite, P. (Ed.). (2024). *Corpora for language learning: Bridging the research-practice divide* (1st ed.). Routledge. <https://doi.org/10.4324/9781003413301>
- Ellis, N. C., & Laporte, N. (2014). Contexts of acquisition: Effects of formal instruction and naturalistic exposure on second language acquisition. In *Tutorials in bilingualism* (pp. 53-83). Psychology Press.
- Francis, W., & Kučera, H. (1964). *Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers*. Brown University.
- Friginal, E. (2013). Developing research report writing skills using corpora. *English for Specific Purposes*, 32(4), 208–220. <https://doi.org/https://doi.org/10.1016/j.esp.2013.06.001>
- Gablasova, D., Brezina, V., & McEnery, T. (2017). Exploring learner language through corpora: Comparing and interpreting corpus frequency information. *Language Learning* 67(S1), 130–154. <https://doi.org/10.1111/lang.12226>
- Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 9–34). Cambridge University Press. <https://doi.org/10.1017/CBO9781139649414.002>
- Gilquin, G., & Granger, S. (2015). Learner language. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 418–436). Cambridge University Press. <https://doi.org/10.1017/CBO9781139764377.024>

- Gilquin, G. (2023). Written learner corpora to inform teaching. In R.R. Jablonkai & E. Csomay (eds) *The Routledge Handbook of Corpora and English Language Teaching and Learning* (pp. 281-295). Routledge.
- Granger, S. (1993). International Corpus of learner English. In Aarts, J., de Haan, P., & Oostdijk, N. (eds.) *English language corpora: Design, analysis and exploitation*, (pp. 57 – 71). Rodopi. https://doi.org/10.1163/9789004653559_007
- Granger, S. (2002). A Bird's-eye review of learner corpus research. In Granger, S., Hung, J., Petch-Tyson, S. (eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3-33). John Benjamins. <https://doi.org/10.1075/llt.6.04gra>
- Granger, S. (2003). The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. In *Tesol Quarterly* 37(3), pp. 538–546. <https://doi.org/10.2307/3588404>
- Granger, S. (2015). The contribution of learner corpora to reference and instructional materials design. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 485-510). Cambridge University Press. <https://doi.org/10.1017/CBO9781139649414.022>
- Granger, S. (2021). Commentary: Have Learner Corpus Research and Second Language Acquisition Finally Met? In B. Le Bruyn & M. Paquot (Eds.), *Learner corpus research meets second language acquisition* (pp. 243–257). Cambridge University Press. <https://doi.org/10.1017/9781108674577.012>
- Granger, S., Dupont, M., Meunier, F., Naets, H., & Paquot, M. (2020). *The International Corpus of Learner English*. Version 3. Presses universitaires de Louvain.
- Granger, S., Gilquin, G., & Meunier, F. (2015). Introduction: learner corpus research – past, present and future. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 1–6). Cambridge University Press. <https://doi.org/10.1017/cbo9781139649414.001>
- Hemchua, S., & Schmitt, N. (2006). An analysis of lexical errors in the English compositions of Thai learners. *Prospect*, 21(3). 3-25.
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.
- Kaya, F. Ö., Uzun, K., & Cangır, H. (2022). Using corpora for language teaching and assessment in L2: A narrative review. *Focus on ELT Journal*, 4(3), 46-62. <https://doi.org/10.14744/felt.2022.4.3.4>
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), 7–36. <https://doi.org/10.1007/s40607-014-0009-9>
- Kučera, H., & Francis, W. (1967). *Computational analysis of present day American English*. Brown University Press. <https://doi.org/10.1002/asi.5090190414>
- Kyle, K. (2016). Measuring syntactic development in L2 writing: Fine grained indices of 97 syntactic complexity and usage-based indices of syntactic sophistication [Doctoral dissertation, Georgia State University]. ScholarWorks @Georgia State University. http://scholarworks.gsu.edu/alesl_diss/35
- Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, 50, 1030–1046. <https://doi.org/10.3758/s13428-017-0924-4>

- Lee, J. J., Bychkovska, T., & Maxwell, J. D. (2019). Breaking the rules? A corpus-based comparison of informal features in L1 and L2 undergraduate student writing. *System*, 80, 143-153. <https://doi.org/10.1016/j.system.2018.11.010>
- Lee, S. (2011). Challenges of using corpora in language teaching and learning: Implications for secondary education. *Linguistic Research*, 28(1), 159–178. <https://doi.org/10.17250/khisli.28.1.201104.009>
- Leech, G. (1981). *Semantics: the study of meaning*. 2nd Ed. Penguin.
- Liao, Y., & Fukuya, Y. J. (2004). Avoidance of phrasal verbs: The case of Chinese learners of English. *Language Learning*, 54(2), 193–226. <https://doi.org/10.1111/j.1467-9922.2004.00254.x>
- Meunier, F. (2020). Introduction to learner Corpus research. In *The Routledge handbook of second language acquisition and corpora* (pp. 23-36). Routledge. <https://doi.org/10.4324/9781351137904-4>
- Moore, J. (2005). *Common mistakes at Proficiency ... and how to avoid them*. Cambridge University Press.
- Murakami, A., & Alexopoulou, T. (2016). L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition*, 38(3), 365-401. <https://doi.org/10.1017/S0272263115000352>
- Myles, F. (2005). Interlanguage corpora and second language acquisition research. *Second Language Research*, 21(4), 373-391. <https://doi.org/10.1191/0267658305sr252oa>
- Myles, F. (2021). Commentary: An SLA perspective on learner corpus research. In B. Le Bruyn & M. Paquot (Eds.), *Learner Corpus Research Meets Second Language Acquisition* (pp. 258–273). Cambridge University Press. <https://doi.org/10.1017/9781108674577.013>
- Nesselhauf, N. (2003). The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied Linguistics*, 24(2), 223–242. <https://doi.org/10.1093/applin/24.2.223>
- O’Keeffe, A., McCarthy, M., & Carter, R. (2007). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge University Press.
- Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. In *Ann Rev Appl Linguist*, 32, 130–149. <https://doi.org/10.1017/S0267190512000098>
- Paquot, M., Larsson, T., Hasselgård, H., Ebeling, S. O., De Meyere, D., Valentin, L., Laso, N. J., Verdaguer, I., & van Vuuren, S. (2022). The varieties of English for specific purposes database (VESPA): Towards a multi-L1 and multi-register learner corpus of disciplinary writing. *Research in Corpus Linguistics*, 10(2), 1–15. <https://doi.org/10.32714/ricl.10.02.02>
- Pérez-Paredes, P. (2022). A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011–2015. In *Computer Assisted Language Learning*, 35(1-2), 36–61. <https://doi.org/10.1080/09588221.2019.1667832>
- Schneider, G. (2023). Detecting and analysing learner difficulties using a learner corpus without error tagging. In K. Harrington & P. Ronan (Eds.), *Demystifying corpus linguistics for English language teaching* (pp. 229–257). Springer International Publishing. https://doi.org/10.1007/978-3-031-11220-1_12
- Selivan, L. (2023). Corpus linguistics and vocabulary teaching. In K. Harrington & P. Ronan (Eds.), *Demystifying corpus linguistics for English language teaching* (pp.

- 139–161). Springer International Publishing. https://doi.org/10.1007/978-3-031-11220-1_8
- Sinclair, J. M. (1990). *Collins COBUILD English grammar*. Collins.
- Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, 97(S1), 77-101.
- Thewissen, J. (2015). *Accuracy across proficiency levels: A learner corpus approach*. Presses universitaires de Louvain.
- Xiao, R. (2009). How can corpora help in language pedagogy. In *Postgraduate Conference in Applied Linguistics, Ningbo, China*.
- Xu, Q. (2016). Application of learner corpora to second language learning and teaching: An overview. In *English Language Teaching*, 9(8), pp. 46–52. Available online at <https://eric.ed.gov/?id=EJ1104573>

Appendix A

Sample Errors and Corrections

FORMAL ERRORS	ERROR	CORRECTION
Formal Misselection	<err type="fm">fastly</err>	<corr type="fm">fast</corr>
	<err type="fm">foreigns</err>	<corr type="fm">foreigners</corr>
Misformation	<err type="mi">copy</err>	<corr type="mi">cheat</corr>
	<err type="mi">traffic</err>	<corr type="mi">traffic jams</corr>
Distortion	<err type="dt">happing</err>	<corr type="dt">happening</corr>
	<err type="dt">congection</err>	<corr type="dt">congestion</corr>
SEMANTIC ERRORS	ERROR	CORRECTION
Confusion of sense relations	<err type="sr">promote</err>	<corr type="sr">encourage</corr>
	<err type="sr">subject</err>	<corr type="sr">issue</corr>
Collocation	<err type="cl">intensive traffic</err>	<corr type="cl">heavy traffic</corr>
	<err type="cl">kill friendship</err>	<corr type="cl">destroy friendship</corr>
Collocation (L1)	<err type="cl">do walking</err>	<corr type="cl">walk</corr>
	<err type="cl">do class</err>	<corr type="cl">attend a class</corr>
Collocation (prep.)	<err type="clp">attend to class</err>	<corr type="clp">attend class</corr>
	<err type="clp">pass on exams</err>	<corr type="clp">pass exams</corr>
Connotation	<err type="ce">affect</err>	<corr type="ce">pollute</corr>
	<err type="ce">help</err>	<corr type="ce">cause</corr>
Style	<err type="sy">licence plate numbers</err>	<corr type="sy">licence plates</corr>
	<err type="sy">kind of</err>	<corr type="sy">partially</corr>

Appendix B

Essay Questions

1. Cars should be banned from city centres to reduce traffic problems in big cities. Do you agree or disagree? To what extent do you agree? Explain your reasons using examples.
2. Universities should adopt a hybrid education model instead of online education. Do you agree or disagree? To what extent do you agree? Explain your reasons with detailed examples.
3. Empathy is considered to be one of the essential personal/social skills in the 21st century. Do you agree or disagree/To what extent do you agree? Explain your reasons using examples.
4. The Internet has caused people to be isolated from their real lives. Do you agree or disagree? To what extent do you agree? Explain your reasons using examples.