Language Resources and Evaluation
https://doi.org/10.1007/s10579-023-09712-w

ORIGINAL PAPER

Analyzing learner language: the case of the Hebrew Learner Essay Corpus

Chen Gafni[1] · **Livnat Herzig Sheinfux**[1] · **Hadar Klunover**[1] · **Anat Bar Siman Tov**[2] · **Anat Prior**[3] · **Shuly Wintner**[1]

Accepted: 21 November 2023

Abstract

We present the Hebrew Learner Essay Corpus (HELEECS): an annotated corpus of Hebrew language argumentative essays authored by prospective higher-education students. The corpus includes essays by two main populations: (1) essays by native speakers of Hebrew, written as part of the psychometric exam that is used to assess their future success in academic studies; (2) essays by non-native speakers of Hebrew, with three different native languages (Arabic, French, and Russian), that were written as part of a language aptitude test. The corpus is uniformly encoded and stored. The non-native essays were annotated with target hypotheses (i.e., hypothesized intended formulations in standard written Hebrew). The corpus is available for research purposes upon request. We describe the corpus and the error correction and annotation schemes used in its analysis. In addition to introducing this new resource, we discuss the challenges of identifying and analyzing non-native language use. Among these challenges are determining whether the language used in a particular utterance is native-like, and determining the target hypothesis when language use is non-native-like. We propose various ways for dealing with these challenges.

Keywords Learner corpora · Hebrew · Non-native language · Crosslinguistic influence · Educational applications

✉ Chen Gafni
chen.gafni@gmail.com

[1] Department of Computer Science, University of Haifa, 3498838 Haifa, Israel

[2] National Institute for Testing & Evaluation, 9139001 Jerusalem, Israel

[3] Department of Learning Disabilities, University of Haifa, 3498838 Haifa, Israel

Published online: 15 May 2024

1 Introduction

Learner corpora—the systematic collection of spoken or written language produced by learners of a language—have been used in research since the late 1980s (De Knop & Meunier, 2015; Granger, 2002; Granger et al., 2015; Tono, 2003). Learner corpora can follow different designs, be of different sizes, involve different language pairs, etc.[1] One paradigm in analyzing learner corpora is the quantitative comparison of categories (words, multi-word expressions, parts of speech, etc.) between learner corpora and native speaker corpora (Gilquin, 2008; Granger, 1996, 2015). This approach, which we follow here, is often called *Contrastive Interlanguage Analysis*. The quantitative analyses range from descriptive comparisons, such as overuse/underuse studies (Durrant & Schmitt, 2009; Gilquin & Paquot, 2008; Hirschmann et al., 2013) to more involved statistical methods, up to modeling (Gries, 2008, 2015; Gries & Deshors, 2015; Vyatkina et al., 2015).

Learner and other non-native language corpora have been instrumental in several tasks, including automatic detection of highly competent non-native writers (Bergsma et al., 2012; Estival et al., 2007; Tomokiyo & Jones, 2001), identification of learners' native language (Bykh & Meurers, 2012; Goldin et al., 2018; Koppel et al., 2005; Tetreault et al., 2013; Tsvetkov et al., 2013), and typology-driven error prediction in learners' language production (Berzak et al., 2015).

In this article, we present the Hebrew Learner Essay Corpus (hereinafter, HELEECS)[2]: an annotated corpus of Hebrew language argumentative essays authored by prospective students in higher-education. The corpus includes both essays by native (or near-native) speakers, written as part of a college entry exam that is used to assess their future success in academic studies; and essays authored by non-native speakers, with three different native languages (Arabic, French, and Russian), written as part of a language aptitude test, also geared towards higher-education admission. The corpus is uniformly encoded and stored. The non-native essays were annotated with target hypotheses (Reznicek et al., 2013), that is, hypothesized intended formulations in standard written Hebrew, whose main goal was to make the texts amenable to automatic processing (morphological and syntactic analysis), thereby guaranteeing uniform representation and processing of the entire dataset.

The current article thus makes two main contributions. The more specific one is the introduction of a new language resource, namely *HELEECS*. More generally, we propose guidelines and recommendations for meaningful linguistic analysis of non-native texts, which take into account the inherent variability of language, with a focus on Hebrew as the target language. The corpus documentation includes guidelines for specific issues in non-native Hebrew, intended to minimize variability between annotators as much as possible. In addition, it includes general guidelines intended to increase the awareness of annotators to the issue of linguistic variability.

[1] For a list of learner corpora, see Learner Corpora around the World; for an extensive bibliography covering learner corpus analyses, see the resources page of the Learner Corpus Association.

[2] This is a revised and much extended version of Gafni et al. (2022), also including some material presented in Nguyen and Wintner (2022).

We demonstrate the utility of the corpus by analyzing many linguistic features of the essays. We were able to attribute some of our findings to known properties of the authors' native languages, thus demonstrating possible effects of linguistic transfer.

The structure of this article is as follows: in Sect. 2, we describe the transcription conventions used in this article. After reviewing some pertinent morphological and orthographic features of Hebrew in Sect. 3, we describe the corpus (Sect. 4) and the process of error correction and annotation (Sect. 5). Section 6 describes two use cases of the corpus. We conclude in Sect. 7 with suggestions for future research.

2 Transcription

In order to properly reflect the errors in learner Hebrew written texts, we use transliteration when discussing examples from the corpus. We opted for a transliteration system rather than a phonetic transcription system, since the intended pronunciation of written words is not always known, especially when the word contains orthographic errors. Table 1 describes the conversion between graphemes (characters or sequences of characters representing a single sound) and transliteration symbols used in this article.[3] The phonetic value of each symbol is also given according to IPA. For some letters, more than one phonetic value is specified according to the letter's pronunciation in given words, by speakers of different ethnolects, or in different contexts. The transliteration system is only used in this article (and the accompanying appendix) for the convenience of readers who are not familiar with the Hebrew script. The corpus itself is written in Hebrew.

Throughout this article, transliterated forms appear between angle brackets and phonetic forms appear between square brackets. Hypothetical phonological representations are enclosed between slashes. When referring to pronunciation, we use a broad phonetic transcription system which is commonly used in Hebrew linguistics. In this system, vowels are transcribed explicitly, unpronounced letters are not transcribed, and pronounced letters are transcribed according to their common pronunciation. Specifically, the following non-IPA conventions are used: [ʔ] = glottal stop, [x] = voiceless velar/uvular/pharyngeal fricative, [š] = voiceless postalveolar fricative, [r] = a rhotic, usually a voiced uvular approximant.

Some transliterated examples in the article contain hyphens that do not exist in the original written texts (e.g., הצעירים <h-cʕyrym̄> 'the-young.PL.M'). These hyphens are morpheme separators added for clarity.

[3] There are various transliteration systems of Hebrew into Latin characters (Gadish, 2012). The system used in this article is closest to Ornan (2017), with several adaptations made in order to accommodate orthographic issues, such as the distinction between word-final and non-word-final letterforms.

Table 1 The transliteration system used in this article

Grapheme	Transliteration	IPA
א	ʔ	ʔ/∅/V
ב	b	b/v
ג	g	g
ג׳	j	ʤ
ד	d	d
ה	h	h/∅/V
ו	w	v/w/V
ז	z	z
ז׳	ž	ʒ
ח	x	x/χ/ħ
ט	ṭ	t
י	y	j/V
כ	k	k/x/χ
ך	k̄	x/χ
ל	l	l
מ	m	m
ם	m̄	m
נ	n	n
ן	n̄	n
ס	s	s
ע	ʕ	ʔ/ʕ/∅
פ	p	p/f
ף	p̄	f
צ	c	ʦ
צ׳	č	ʧ
ץ	c̄	ʦ
ץ׳	c̄̆	ʧ
ק	q	k
ר	r	ʁ
ש	š	ʃ/s
ת	t	t
״	"	∅
׳	׳	∅

Notes

1. A double over-line represents word-final letterform (see Sect. 3.1).

2. The empty set symbol ∅ represents an unpronounced letter.

3. V = some vowel. The four *matres lectionis* א, ה, ו, י can represent both consonants and vowels. Each *mater lectionis* can represent several of the five vowels of Hebrew. More specifically, א and ה usually represent the vowel [a] or [e], ו represents [o] or [u], and י usually represents [i], but sometimes [e].

4. The Gershayim symbol (double quotation) is often used in Hebrew acronyms and abbreviations (Jacobs et al., 2020). Though it has no phonetic value, it was kept in the transliteration in this article (e.g., חו״ל < xw"l > 'abroad' is an acronym of חוץ לארץ < xw c̄ l-ʔr c̄ >, lit.: 'out to the country'). The Geresh symbol (single quotation) can be

Table 1 (continued)

	used as a phonetic modifier (e.g., ג<g> →ג׳< j >), in which case it does not appear in the transliterated form. A Geresh can also be used as an abbreviation marker, in which case it is retained in the transliterated form (e.g., נק׳ <nq′> is an abbreviation of נקודה <nqwdh> 'dot, point').

3 Linguistic properties of Hebrew (with implications to learning)

L2-Hebrew learners face many challenges on their way to becoming proficient users. Among these challenges are the abjad orthography and complex morphology of Hebrew (Fabri et al., 2014).

3.1 Orthography

Hebrew is a Semitic language, like Arabic and Amharic. Similarly to Arabic, Hebrew is written right-to-left, and its orthographic system is consonant-based (i.e., an *abjad system*). Vowels do not have dedicated letters. They are represented inconsistently and incompletely by the dual-function letters {א, ה, ו, י} {<ʔ, h, w, y>}, which can represent both consonants and vowels.[4] Each of these letters denotes at least two vowels, and the formal standard use of these letters as vowels is subject to various conventions, usually morphological. For example, the endonym of the Hebrew language ['ivrit] contains two [i] vowels. In the standard written version, עברית, the second [i] is represented by the vowel letter Yod י, while the first vowel is not represented in the script. As a result of the underrepresentation of vowels in the script, Hebrew is characterized by a vast amount of homographs (Bentin & Frost, 1987; Share & Bar-On, 2018).

Another aspect of homography in Hebrew is due to the existence of several letters that can each represent multiple consonants. For example, the letter Kaf (כ) can represent the consonants [k] and [x]. This duality of Kaf is due to a context-dependent historical phonological process of spirantization, by which the plosive consonant /k/ is realized as a fricative consonant [x] after a vowel. Spirantization is most conspicuous in verbal paradigms. For instance, כ represents [k] in ישכב [yiškav] 'lie down.3SG.M.FUT', but [x] in שכב [šaxav] 'lie down.3SG.M.PST'. In addition to homography, Hebrew also has several letters representing the same sound (homophonic letters). For instance, the voiceless velar stop [k] is represented by two letters: Kaf (כ) and Qof (ק). Thus, Hebrew learners face multiple challenges in learning the form-sound mapping of the Hebrew orthography (e.g., learning when [k] is represented by כ and when by ק, as well as learning when כ is pronounced as [k] and when as [x]).

[4] Hebrew vowels may also be represented by diacritical marks called *nikkud* placed above, below, or inside letters. However, Hebrew texts usually appear without nikkud (i.e., *unpointed*, *undotted*, or *unvoweled* form). Texts that do contain nikkud (i.e., *pointed*, *dotted*, or *voweled* form) appear mainly in children's books, holy scripts, and poetry (Ben-Dror et al., 1995). Beginning readers rely on pointed texts to learn the basics of Hebrew, and the transition to reading unpointed texts poses a challenge to them in their way of becoming expert readers (Share & Bar-On, 2018). The essays in the corpus are unpointed.

The visual form of letters introduces another level of complication to the Hebrew orthography. First, five letters of the Hebrew alphabet have a different form when appearing in word final (ך, ם, ן, ף, ץ) vs. word non-final (כ, מ, נ, פ, צ) positions. Second, some pairs of Hebrew letters have similar visual appearance. For instance, the letters Heh ה<h>and Heth ח<x>consist of a horizontal line placed above two parallel vertical lines, and the only difference between the letters is that in Heth, the left vertical line is connected to the horizontal line, while in Heh, there is a small gap between them. Three letters (ו, ן, י) only differ in the length of their vertical lines. Yet another letterform-related complication arises from the existence of two parallel Hebrew script systems. *Block* (*square*, or "*print*") script is the main type of script used in printed (or typed) texts (such as this article), while *cursive* ("*handwritten*") script is the main type used in handwritten texts. Some letters have a similar form in both script systems (e.g., the block variant of the letter Heh is ה and its cursive variant is ה), while other letters have different appearances across scripts (e.g., the block variant of the letter Mem is מ and its cursive variant is מ). It just so happens that there are some pairs of letters that look similar in cursive script (though not in block script). For example, the cursive variants of the letters Gimel (ג,<g>) and Zayin (ז,<z>) are mirror images of each other.

Overall, visual similarities across letters and context-dependent variation of letterform may be confusing for learners. Indeed, HELEECS contains many instances of errors that appear to reflect such confusions. For example, four essays in HELEECS contain the nonword גמן (<gm$\bar{n}$>), which is likely a deformation of the word זמן (<zm$\bar{n}$> 'time'). Since the essays in HELEECS were originally handwritten, a probable explanation for the error is confusion between the cursive forms ג and ז.

3.2 Morphology and morpho-orthography

Like other members of the Semitic language family, Hebrew has a rich morphological system that is largely non-linear (or *non-concatenative*). All verbs and many nouns and adjectives are formed by interleaving a consonantal root within a morphological pattern. The consonantal root is made up of (usually three) consonants. It is assumed to represent an abstract concept that frequently carries the primary meaning component of the word and serves as a lexical access unit (Frost et al., 2000; Gafni et al., 2019; Prior & Markus, 2014; Ravid, 2020; Ravid & Malenky, 2001; Schwarzwald, 2002; Shimron, 2003). The morphological pattern is composed of several vowels and, occasionally, some consonants, in fixed positions, with open slots into which the root's consonants can be inserted. It represents a combination of properties such as lexical category, tense, gender, aspect, and so on. For example, the consonantal root נ-ס-ח<X-S-N>stands for the concept of immunity, or strength.[5] It can be incorporated in various patterns to create words such as חיסון

[5] Uppercase letters in transcribed forms represent root consonants. When shown in isolation, roots are represented by their transliterated form (e.g., <X-S-N>) for reasons of convenience. In such forms, root letters are separated by hyphens.

([XiSuN], noun: 'vaccine'), התחסנו ([hitXaSNu], verb: 'get vaccinated.3PL.PST') and חסין ([XaSiN], adjective: 'immune.SG.M, resistant.SG.M').[6]

Mastering the morphological system of Hebrew can be rather challenging for any L2-Hebrew learner. Beginner learners with no knowledge of other Semitic languages need to learn to process non-linear morphology and acquire the independent representations of roots and patterns (e.g., Norman et al., 2016). By contrast, native speakers of another Semitic language (e.g., Arabic) are already equipped with the skills required for processing non-linear morphology. However, such speakers need to learn to suppress the knowledge of their L1, which might cause them to transfer both roots and patterns from their L1 into Hebrew (Abu Baker, 2016).

The Hebrew morphological system is challenging mainly because of the large number of possible patterns, the similarities among patterns, and the fact that many of the patterns and roots are semantically ambiguous or opaque. For example, some pairs of patterns differ in the position of a single letter, which can cause changes in the lexical category, gender, and tense, among other things. Moreover, such a minor orthographic change can also result in the generation of a non-existing word or an existing, but semantically unrelated, word. For example, consider the patterns CyCC and CCyC.[7] The only visual difference between their written forms is the position of the letter י (<y>). When a consonantal root is embedded in these patterns, the resulting pair of words can have any semantic relation. In the case of the root מ-ה-ר <M-H-R>, the obtained words are semantically related: a verb (CyCC: מיהר [MiHeR] 'rush.3SG.M.PST') and an adjective (CCyC: מהיר [MaHiR] 'fast.SG.M'). In the case of the root ח-ז-ר <X-Z-R>, the obtained words are semantically unrelated: a verb (CyCC: חיזר [XiZeR] 'court.3SG.M.PST') and a noun (CCyC: חזיר [XaZiR] 'pig.M'). Moreover, the root ט-ג-נ <Ṭ-G-N> produces an existing verb in the CyCC pattern (טיגן [TiGeN] 'fry.3SG.M.PST'), but a non-existing word in the CCyC pattern (*טגין [TaGiN]). In the latter case, Hebrew speakers might spontaneously assign the meaning 'something that can be fried' to טגין, indicating the productivity of the root-and-pattern system. This is less likely to be so in the case of חזיר, since the wordform already exists and has an unrelated meaning.

Hebrew also has an additional morpho-orthographic property that might be confusing for learners and also difficult for automatic parsers. Hebrew has seven function clitics consisting of a single letter that is prefixed in the script to the hosting word (Fabri et al., 2014). These include the definite article ה <h-> 'the', the coordinating conjunction ו <w-> 'and', the subordinator ש <š-> 'that', and the prepositions ב <b-> 'in', כ <k-> 'as', ל <l-> 'to' and מ <m-> 'from'. Other function words appear standalone in the script. One challenge with these cliticized letters stems from the fact that they can also be a part of a lexical unit. As a result, there are many morphologically ambiguous orthographic words in Hebrew (e.g., Share & Bar-On,

[6] In the context of verbs, our working definition of morphological patterns covers not only the basic forms (e.g., the form [hitC$_1$aC$_2$eC$_3$], which is the third person, singular, masculine form of Binyan Hitpa'el), but also forms representing different verb conjugations (e.g., the form [titC$_1$aC$_2$C$_3$i], which is the second person, singular, feminine future form of Binyan Hitpa'el). C$_1$, C$_2$, and C$_3$ represent root consonants.

[7] Uppercase C represents any root consonant letter.

Table 2 Readings of the morphologically ambiguous לבן

	Reading 1		Reading 2	
a. Non-clitic ל:	[laván]	'white.SG.M'	[lében]	'Leben (dairy)'
b. Clitic ל:	[la-bén]	'to the boy'	[le-bén]	'to a boy'/ 'to Ben'

Note: acute accent (é) indicates a stressed vowel

2018). For example, in the word לבן the first letter ל can be either a part of the lexical unit (Table 2, row a) or a clitic element (Table 2, row b). The exact reading is context-dependent.

Cliticized letters can also be combined in a chain of prefixes, thus making their identification more difficult. For example, ושהתמונה <w-š-h-tmwnh> 'and that the picture' is composed of a noun base with three cliticized prefixes.

Overall, the morpho-orthographic properties of Hebrew make it orthographically dense. That is, the number of orthographic neighbors of a given word is much higher than in many other languages (e.g., Frost, 2012).[8] In particular, letter-transposition neighbors are relatively common in Hebrew (e.g., תועלת <twʕlt> 'benefit' – תולעת <twlʕt> 'worm'; חלב <xlb> 'milk' – חבל <xbl> 'rope' / 'a shame'; חזר <xzr> 'return.3SG.M.PST' – חרז <xrz> 'rhyme.3SG.M.PST'). As a result of the high orthographic density of Hebrew, orthographic and morphological errors often yield, by pure chance, another existing word, which is not necessarily semantically related to the intended word.

3.3 Morpho-syntax

The syntactic and morpho-syntactic properties of Hebrew also deserve some attention in the context of learner language. Hebrew has grammatical gender (masculine and feminine) which is marked on nouns, verbs, adjectives, and more. Since grammatical gender of inanimate objects is arbitrary, second language learners typically struggle with acquiring the gender system of their L2, irrespective of their L1 (e.g., Sabourin et al., 2006). This struggle can be expressed through the difficulty of choosing the correct verbal and adjectival forms to maintain agreement with the gender of the modified noun. For example, in French *problème* 'problem' is masculine, while the Hebrew equivalent בעיה [be'aya] is feminine. An essay in HELEECS by an L1 French author contained the ungrammatical phrase הבעיה האחרון [ha-be'aya ha-axaron] 'the final.M problem(F)', in which the gender of the adjective seems to reflect the gender of the noun in French, but not in Hebrew.[9]

In addition to Hebrew-specific properties, it should be noted that there are various aspects of language that learners typically struggle with, regardless of the language

[8] An (immediate) orthographic neighbor is a word created from another word (e.g., *trail*) by a single orthographic change, including an insertion (e.g., *trails*), deletion (e.g., *rail*), substitution of a single letter (e.g., *train*), or transposition of two letters (e.g., *trial*).

[9] In the equivalent French expression, *le dernier problème* 'the.SG.M final problem(M)', agreement with the noun is achieved through the definite article rather than the adjective.

in question. These include correct use of prepositions, conjunctions, determiners, and lexical items. These properties are universally difficult for L2 learners due to the arbitrariness of form-meaning mappings across languages. For example, each of the following phrases contains a different preposition in English: *at home*, *on Monday*, *in January*. By contrast, the equivalent Hebrew phrases use the same preposition, ב<b->. Conversely, phrases that contain the same preposition in English are equivalent to Hebrew phrases that contain different prepositions. For example, the following English phrases contain the preposition *for*: *I did it for Dan* and *I am waiting for Dan*. The equivalent Hebrew phrases contain the prepositions ל<l-> and בשביל<bšbyl>, respectively. Overall, there is no simple mapping between prepositions in the two languages (see also Hermet & Désilets, 2009).

4 The corpus

4.1 The essays

The corpus[10] includes 3000 argumentative essays authored by non-native speakers of Hebrew, distributed equally over three native languages (L1s): Arabic, French, and Russian. In addition, it includes 1000 essays in Hebrew authored by native speakers. The essays in both collections were written by examinees as part of the admission process to higher-education institutions in Israel. The essays by Hebrew native speakers were written as part of the Psychometric Test (National Institute for Testing & Evaluation, 2012), a general test required for admission by most higher-education institutions in Israel (the test is administered in several languages). The essays by non-native speakers were collected as part of the *YAEL* test (National Institute for Testing & Evaluation, 1986): a Hebrew proficiency test required for examinees who chose to sit the Psychometric Test in a language other than Hebrew. Both tests are administered by the Israeli National Institute for Testing & Evaluation (*NITE*), from which we obtained the essays.

In the absence of direct information, the authors' native languages were determined based on the language in which they chose to take the Psychometric Test. Those authors who chose to take the test in Hebrew are viewed as native speakers because the choice suggests a high proficiency in Hebrew. Similarly, authors who chose to take the test in Arabic, French, or Russian are considered native speakers of those languages.

Essays in the YAEL sub-corpus were written in response to one of nine prompts, while essays in the Psychometric sub-corpus were written about one of two topics (the prompts for the two sub-corpora differ). The Psychometric (native) essays were collected in 2012 (topic 1) and 2017 (topic 2). The YAEL (non-native) essays were collected between the years 2011–2020. The conditions and requirements of

[10] The corpus is available for research purposes upon request and subject to signing a license agreement form. Additional information about the corpus is provided as data statements (Bender & Friedman, 2018) in the accompanying datasheet (Gebru et al., 2020). The corpus and accompanying documents can be found at https://github.com/HaifaCLG/HebrewEssayCorpus.

the tests also differed between the two groups: the allotted time for essay writing was 15 min in the YAEL test and 30 min in the Psychometric Test. In addition, there was a specific length requirement for each test: 10–15 lines in YAEL and 25 lines in the Psychometric Test.

4.2 Metadata

The only available metadata for the native speaker essays is the essay score, in the range 1–6 (mean: 3.67). The essays included in this sub-corpus were selected randomly with no exclusion criteria. Essays in the non-native sub-corpus are accompanied by the following metadata (some pieces of information are unavailable for some essays):

- **Author's L1**: Arabic, French, or Russian.
- **Gender**: Male, Female, Unspecified.
- **Age**: 13–50 (mean: 21, SD: 4).
- **Year of exam**: 2011–2020.
- **Prompt**: 1–9, representing the topic of the essay (the explicit prompts are confidential).
- **Essay score**: the range of scores for essays included in the corpus is 17–28 (mean: 20.7, SD: 2.4).[11] These scores were assigned by two professional NITE raters.
- **Scores of components of essay evaluation**: these include (i) Content, (ii) Organization, (iii) Linguistic Richness, and (iv) Linguistic Precision. The range of each component grade is 1–7.
- **Total Psychometric score**: the scores of the Psychometric Test have a normal distribution in the range 200–800 with a mean of 550. The Psychometric scores of candidates whose essays are included in our corpus were in the range 279–778 (mean: 540, SD: 97).
- **Scores of Psychometric components**: (i) Verbal Reasoning, (ii) Quantitative Reasoning, and (iii) English. The range of each component is 50–150.
- **Parental education (for each parent)**: no education, primary, partial secondary, full secondary, partial tertiary, academic degrees: bachelor, master, doctoral.
- **Family income**: six levels ranging from very low to very high, plus unspecified income.

Table 3 summarizes the mean number of sentences and tokens per essay in each of the three L1s. The mean number of sentences per essay in the native sub-corpus was 15.2 (SD: 5.3), and the mean number of tokens was 329 (SD: 81). These numbers are considerably higher than in the non-native essays. However, the length differences across the two sub-corpora are likely due to the test requirements (see Sect. 4.1).

[11] The full range of scores in the YAEL test is 4–28.

Table 3 Mean numbers of sentences and tokens per essay across L1s in the non-native corpus. Numbers in parentheses denote standard deviation

	Arabic	French	Russian
Sentences	6.1 (2.6)	9.0 (2.8)	8.9 (2.7)
Tokens	143 (28)	142 (29)	138 (27)

Note that the number of sentences in essays authored by L1 Arabic speakers was considerably lower than in the other two L1s, although the total number of tokens was similar across the three L1s. We discuss this observation in Sect. 6

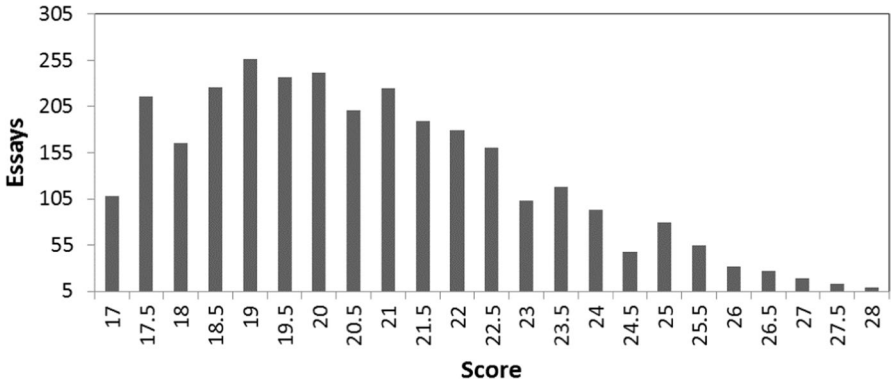

Fig. 1 Distribution of essays by score

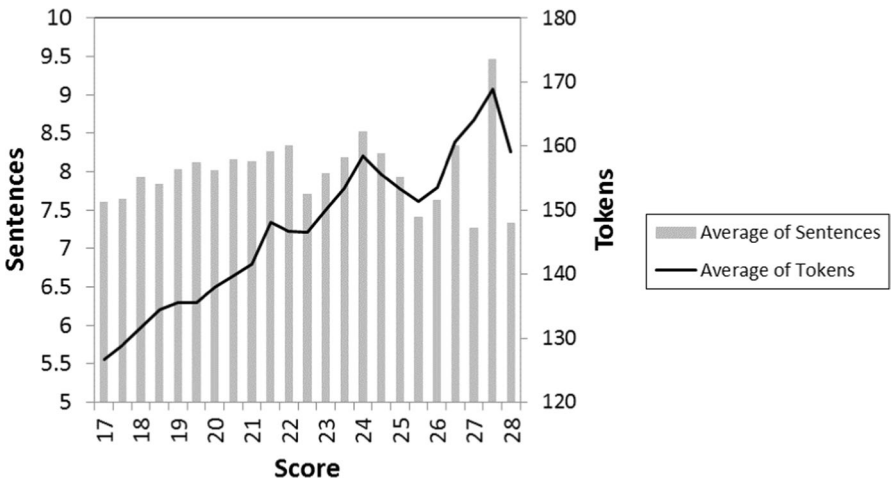

Fig. 2 Mean numbers of tokens and sentences per essay across YAEL test scores

Figure 1 shows the distribution of essays in the non-native sub-corpus by score. The distribution is evidently normal, but its lower (left) part is truncated by design: we requested only essays above a certain score, because the level of Hebrew in the

lowest-scored essays was too low to allow effective and informative analysis. Scores can be non-integral because they represent the average of the two human-assigned scores.

Figure 2 depicts the mean number of sentences (represented as bars) and tokens (represented as a curve) per essay across the non-native test scores. The number of tokens is positively correlated with the test score (Pearson's $r = 0.29$, $p < 0.001$), while the number of sentences is not (Pearson's $r = 0.03$, $p = 0.14$).

4.3 Processing

The essays, originally hand-written, were transcribed by NITE and stored in text files. The typist was instructed to type the essays precisely as written, including spelling and tokenization errors. Unreadable words were replaced by a tilde. The typed essays were then sample inspected by a supervisor. The order of sentences in each essay was scrambled before the files were delivered to us to preserve author privacy. We tokenized the entire dataset using Child Phonology Analyzer (Gafni, 2015). The tokenized essays were stored in a tabulated format to facilitate error correction and annotation. Table 4 illustrates the processed representation of sentence (1) and its revision (1′).

(1) *וככה **ה צעירים יכלו** ללמוד בדיוק אחרי **לסיום** בית ספר

w-kkh	**h**	**cʕyrym̄**	**yklw**	llmwd	bdywq	ʔxry	**l-sywm̄**
and-this way	**the**	**young.PL.M**	**can.3PL.PST**	study.INF	exactly	after	**to-end.CONSTR**

byt	spr
house.CONSTR	book

(1′) וככה **הצעירים יוכלו** ללמוד בדיוק אחרי **סיום** בית ספר

w-kkh	**h-cʕyrym̄**	**ywklw**	llmwd	bdywq	ʔxry	**sywm̄**
and-this way	**the-young.PL.M**	**can.3PL.FUT**	study.INF	exactly	after	**end.CONSTR**

byt	spr
house.CONSTR	book

'And this way the young could study right after school graduation'

Tokens of the original text were stored in a column labelled "Token", while revised tokens were stored in a column labelled "TH1" (standing for "Target hypothesis1"). Deletion, insertion, splitting, or merging of words was indicated by the insertion of a "&&" dummy token at the relevant position to maintain the alignment between the texts (e.g., a dummy token was inserted in row 2 in Table 4 to maintain alignment between the revised token הצעירים < h-cʕyrym̄ > 'the-young.PL.M' and the corresponding split token in the original text ה צעירים < h cʕyrym̄ > 'the young.PL.M').

Table 4 A processed text

	Token	TH1
1	w-kkh	w-kkh
2	**h**	**&&**
3	**cʕyrym̄**	**h-cʕyrym̄**
4	**yklw**	**ywklw**
5	llmwd	llmwd
6	bdywq	bdywq
7	ʔxry	ʔxry
8	**l-sywm̄**	**sywm̄**
9	byt	byt
10	spr	spr

The tokenization scheme used in the processing of the corpus was based on orthographic words rather than on syntactic words, as used in some parsers like YAP (More et al., 2019) and UD HTB (Zeldes et al., 2022). While such morpho-syntactic parsers are valuable for analyzing Hebrew texts in general, it is important to note that they were trained on normative native Hebrew and are therefore less suitable for parsing learner language, which may contain complex mixtures of errors on all levels of linguistic analysis. In addition, as noted in Sect. 3.2, due to the morpho-orthographic properties of Hebrew, even texts in normative Hebrew can be challenging for morpho-syntactic parsers.

5 Annotation and target hypotheses

We reviewed the essays and annotated various types of errors. Most essays were reviewed by one annotator, except for 54 essays that were reviewed by two annotators to assess inter-annotator agreement (see Sect. 5.3). All annotators were native speakers of Hebrew, with an undergraduate or a graduate degree in linguistics. The remainder of this section details our annotation scheme. Overall, we annotated 1013 essays out of the 3000 non-native ones. Table 5 specifies the number of annotated essays, sentences, and tokens per L1. The distribution of annotated essays over test scores is shown in Fig. 3 (this distribution is a subset of the one shown in Fig. 1, which includes non-annotated essays as well).

Our annotations consist of three distinct pieces of information. First, there's the indication that a sentence is ill-formed; this is done by marking tokens in the sentence that cause deviation from standard language. Second, we propose target hypotheses to replace these marked tokens (see Sect. 5.1). Finally, we also offer interpretations pertaining to the presumed cause of these errors, formulated as a basic classification of the errors by type/cause (see Sect. 5.2).

Table 5 Statistics of annotated non-native essays per L1

L1	Essays	Sentences	Tokens
Arabic	342	2023	50,304
French	338	2989	48,893
Russian	333	2993	47,213
Total	1013	8005	146,410

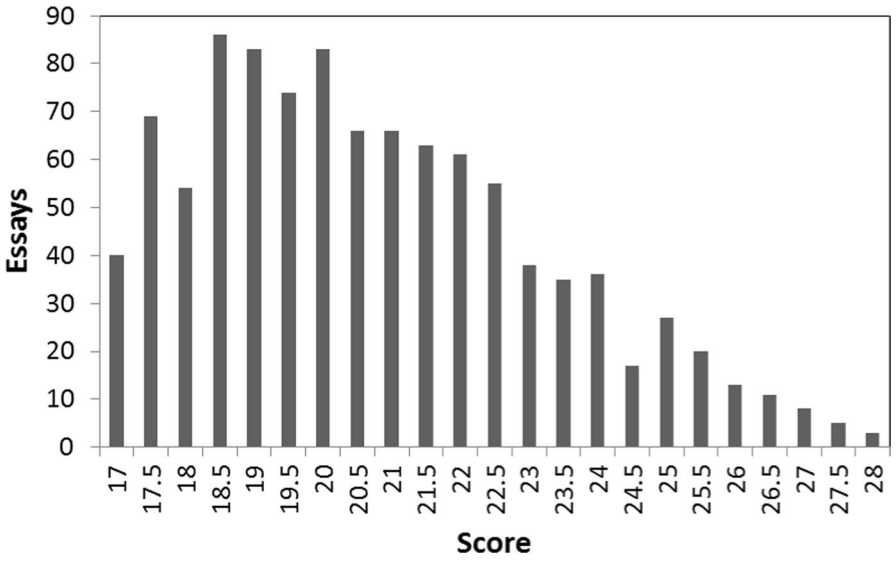

Fig. 3 Number of annotated non-native essays per test score

5.1 Principles of the target hypothesis

When correcting a non-native text, it is sometimes assumed that the language used deviates in some way from "typical", or "standard" native language use, and that the author's intended meaning can be recovered and reconstructed according to the norms of the target language. In reality, this is not a straightforward matter. First, the notion of "standard" native language is elusive: native speakers vary greatly in their use of language, and more often than not avoid adhering to prescriptive language norms (Dąbrowska, 2018). Second, it is impossible to construct with certainty an utterance in native-like language that would retain the author's intended meaning, simply because this meaning is not part of the text and is thus unknown.

Therefore, generating an equivalent "native-like" version of a non-native text is a difficult, ill-defined task. Instead, we adopt an approach that minimally modifies the non-native texts by associating some (ill-formed) constructions with a *target hypothesis* (Reznicek et al., 2013). Our goal is to introduce a minimal number of changes in an input sentence in order to obtain a grammatically correct and contextually appropriate utterance in the target language that would make the resulting utterance

amenable to automatic language processing tools, such as a morphological analyzer and a parser.

In this project, we adopted a broad interpretation of the term "grammar", to potentially cover all levels of linguistic analysis on which native and non-native language use can be distinguished, including orthography, morphology, syntax, semantics, and discourse. This decision was motivated by the theoretical conception of language as a whole, but also by the properties of Hebrew that make it difficult to tease apart different levels of analysis (see Sect. 3).

With this notion of grammaticality in mind, annotators were guided to rely on their intuitions as native speakers of Hebrew, as well as on their experience as linguists, when determining whether the text is native-like and, if not, to induce minimal modifications to make it native-like. As noted above, native language use is inherently variable and, thus, any evaluation and adaptation of texts that is based on speakers' intuitions is bound to yield variable results. Consequently, the annotation process cannot be entirely consistent across (and even within) annotators. Yet, we formulated elaborate guidelines in an attempt to minimize inter-annotator variability as much as possible, and introduced means to include alternative interpretations in the annotations, thus recognizing the inherent variability in language use. In the following sub-sections, we describe several general principles that guided annotators regarding whether or not a fragment of text should be revised, and if so, how to make the most conservative revision.

5.1.1 The grammaticality principle

Annotators were guided to correct any text fragment that deviated markedly from typical native language use, provided that there was a clear grammatically-correct alternative. Moreover, annotators were guided not to dwell on the text trying to guess the intended meaning, but to follow their initial intuition as much as possible.[12] For example, consider the following sentence:

(2) *הוא צריך להשיג **משהו** רוצה

hw?	cryk̄	lhšyg	**mšhw**	rwch
he	need.SG.M.PRS	achieve.INF	**something**	want.SG.M.PRS

*'He needs to achieve **something** wants'

Sentence (2) is ungrammatical. The most conservative interpretation would be to treat משהו < mšhw > [mašehu] 'something' as a morpho-orthographic error, an incorrect merging of the words מה שהוא < mh š-hw? > [ma še-hu] 'what that-he'. The hypothesized target sentence is then:

[12] A common observation among linguists is that engaging in grammaticality analysis for an extended period of time can affect linguistic intuitions and reduce the speaker's confidence in them. This is sometimes referred to as *scanting out* (e.g., Schütze, 2016: 113) or as *syntactic satiation* (Sprouse, 2009).

(2′) הוא צריך להשיג **מה שהוא** רוצה

hw?	cryk̄	lhšyg	**mh**	**š-hw?**	rwch
he	need.SG.M.PRS	achieve.INF	**what**	**that-he**	want.SG.M.PRS

'He needs to achieve **what he** wants'

This is considered a conservative interpretation since it assumes a simple cause for the error: the fact that both phrases are pronounced identically in speech. Additional examples of errors on various levels of linguistic analysis, as well as the treatment of these errors, are provided in Sect. 5.2.

5.1.2 The cooperative principle

In the spirit of the Gricean *cooperative principle* (Grice, 1989), the sensible author is likely to make sensible utterances. In the current framework, a sensible utterance is one that is acceptable on all levels of linguistic analysis by the standards of native speakers (as assumed by the annotators). Under the cooperative principle, we modify sentences that are syntactically and morphologically valid, but inappropriate in the given context (in contrast to the grammaticality principle, which applies to sentences that are unacceptable in any context).[13] The assumption underlying this principle is that the author likely made an error (e.g., orthographic, morphological) rather than intentionally wrote a sentence that does not make sense.

This principle has become a major issue due to the orthographic and morphological structure of Hebrew, where small errors can generate existing but semantically-unrelated words by pure chance (see Sect. 3.2), whereas errors of similar nature typically generate nonwords in other languages. When an error generates a non-existing word, it is easier to agree that the nonword should be corrected. But given the above considerations, we claim the same should also apply when the error generates an existing word (see examples below).

The formal guideline that follows from the cooperative principle is: given a syntactically and morphologically valid sentence that does not make sense – if the sentence can be made sensible via small orthographic/morphological corrections, revising the sentence should be preferred over retaining the original sentence. Orthographic corrections include transposition, insertion, deletion, or substitution of a letter with a phonetically/visually similar letter. Morphological corrections typically involve a change of affix or non-linear pattern (*Binyan* for verbs, *Mishkal* for nouns and adjectives) while retaining the consonantal root. The hallmark of cases that are typically corrected under this principle is a small edit distance between the original and revised token but a large semantic distance (the words belong to different semantic fields). The following example illustrates the application of the cooperative principle:

[13] In HELEECS, the sentences are not given in their original order. Thus, the immediate context of any given sentence is essentially unknown. However, examining the entire essay, we can determine at least a "thematic" context, against which the appropriateness of sentences can be evaluated to some degree. Occasionally, we encountered sentences that were extremely unlikely and could be judged inappropriate even without context (or, with a "zero" context).

(3) זה יבוא רק **לתולעת** המשפחה עצמה

zh	ybwʔ	rq	**l-twlʕt**	h-mšpxh	ʕcmh
it	come.3SG.M.FUT	only	**to-worm.CONSTR**	the-family	herself

'It will come only to the **worm** of the family itself'

Sentence (3) is syntactically correct, but does not make sense in the context in which it appeared (e.g., worms are not mentioned anywhere else in the essay). A plausible explanation for this sentence is a letter transposition error: לתולעת 'to the worm of' should probably have been לתועלת 'to the benefit of'. We annotate this as a spelling error and introduce a correction. The hypothesized target sentence is:

(3′) זה יבוא רק **לתועלת** המשפחה עצמה

zh	ybwʔ	rq	**l-twʕlt**	h-mšpxh	ʕcmh
it	come.3SG.M.FUT	only	**to-benefit.CONSTR**	the-family	herself

'It will be (lit.: come) only to the **benefit** of the family itself'

5.1.3 Minimal editing

The grammaticality and cooperative principles focus mainly on the justification for revising the text. The *faithfulness principle* provides general guidelines for how the revision should proceed. According to this principle, the revised text should be as close as possible in meaning and form (i.e., be faithful) to the original text. In other words, annotators were instructed to keep the editing as *minimal* and local as possible and to avoid rewriting the text extensively to make it sound "better". In practice, if there are several more-or-less equivalent ways of revising the text to make it more native-like, annotators should opt for the option that involves fewer changes, in terms of tokenization and the number of altered words. For instance, sentence (4) is clearly missing a preposition before מחשב <mxšb> 'computer', but there are several suitable alternatives, including ב <b-> 'in', מ <m-> 'from', and באמצעות <bʔmcʕwt> 'using'. In this case, the first two alternatives are preferable, since the prepositions ב and מ are used as clitics and, therefore, do not affect tokenization. This is demonstrated in (4′). By contrast, adding the stand-alone preposition באמצעות is dispreferred, since it increases the number of words in the text (see 4″).

(4) *היום אפשר לקרוא מה קורה בסין **מחשב** שנמצא בפריז

hywm̄	ʔpšr	lqrwʔ	mh	qwrh	b-syn̄	**mxšb**
today	possible	read.INF	what	happen.SG.M.PRS	in-China	**computer**

š-nmcʔ	b-pryz
that-situated.SG.M.PRS	in-Paris

*'Today it is possible to read what happens in China **a computer** located in Paris'

(4′) היום אפשר לקרוא מה קורה בסין **ממחשב** שנמצא בפריז

hywm̄	ʔpšr	lqrwʔ	mh	qwrh	b-syn̄	**m-mxšb**
today	possible	read.INF	what	happen.SG.M.PRS	in-China	**from-computer**

š-nmcʔ	b-pryz
that-situated.SG.M.PRS	in-Paris

'Today it is possible to read what happens in China **from a computer** located in Paris'

(4″) היום אפשר לקרוא מה קורה בסין **באמצעות מחשב** שנמצא בפריז

hywm̄	ʔpšr	lqrwʔ	mh	qwrh	b-syn̄	**bʔmcʕwt**	**mxšb**
today	possible	read.INF	what	happen.SG.M.PRS	in-China	**using**	**computer**

š-nmcʔ	b-pryz
that-situated.SG.M.PRS	in-Paris

'Today it is possible to read what happens in China **using a computer** located in Paris'

5.1.4 Information maximization

A second corollary of the faithfulness principle is the "information maximization" principle. According to this principle, a revised text should retain the maximal amount of information contained in the original text, and add as little information as possible. We assume the following information content hierarchies:

- Content words > function words
- Lexical morphemes > grammatical morphemes

Lexical morphemes include consonantal roots in Semitic languages and monomorphemic content words. Grammatical morphemes include affixes as well as nonlinear morphological patterns in Semitic languages.

In practice, the information maximization principle states that changing lower-order elements on the information content hierarchies is preferred to changing higher-order elements. When two alternative corrections are possible, we implement the one requiring minimal assumptions and minimal modifications of the original text. The following example illustrates this principle.

(5) *יש הורים שלהם אין מספיק כסף כדי **להספיק** להם את כל צרכיהם

yš	hwrym̄	š-lhm̄	ʔyn̄	mspyq	ksp̄	kdy
exist	parents	that-to.them	there is no	sufficient	money	for

lhspyq	lhm̄	ʔt	kl	crkyhm̄
suffice.INF	to.them	ACC	all	needs.POSS.3PL.M

*'There are parents who don't have enough money to **suffice** them all their needs

(5) is ungrammatical due to a mismatch between the verb and complements. The verb להספיק < lhspyq > 'suffice' is assigned two complements here: < lhm̄̄ > 'to them' and < ʔt kl crkyhm̄̄ > 'all their needs'. Of the two complements, only the first

fits into the argument structure of the verb.[14] However, omitting the second complement will lead to a loss of information, and the resulting phrase would still be ungrammatical (or at least odd):

(5′) יש הורים שלהם אין מספיק כסף כדי **להספיק** להם[???]

yš	hwrym̄	š-lhm̄	ʔyn̄	mspyq	ksp̄	kdy	**lhspyq**	lhm̄
exist	parents	that-to.them	there is no	sufficient	money	for	**suffice.INF**	to.them

[???]'There are parents who don't have enough money to **suffice** for them'

The more plausible correction involves changing the verb להספיק < lhspyq > (pronounced: [lehaSPiK]) to a verb of the same root in a different Binyan (verb pattern): לספק < lspq > (pronounced: [leSaPeK]) 'to provide'. The revised verb is compatible with the argument structure of the original sentence. Thus, no information is lost in the revised sentence and the correction requires a single morphological change. The hypothesized target phrase is:

(5″) יש הורים שלהם אין מספיק כסף כדי **לספק** להם את כל צרכיהם

yš	hwrym̄	š-lhm̄	ʔyn̄	mspyq	ksp̄	kdy	**lspq**
exist	parents	that-to.them	there is no	sufficient	money	for	**provide.INF**

lhm̄	ʔt	kl	crkyhm̄
to.them	ACC	all	needs.POSS.3PL.M

'There are parents who don't have enough money to **provide** them all their needs'

Alternatively, one could opt for replacing the verb in (5) with a semantically similar verb from another root, such as לתת < ltt > 'to give', as in (5‴). However, (5‴) involves a change in a lexical morpheme (a root) plus a change in a grammatical morpheme (a morphological pattern), which is less conservative than a change in a grammatical morpheme alone, as in (5″). Therefore (5″) is preferred to (5‴).

(5‴) יש הורים שלהם אין מספיק כסף כדי **לתת** להם את כל צרכיהם

yš	hwrym̄	š-lhm̄	ʔyn̄	mspyq	ksp̄	kdy
exist	parents	that-to.them	there is no	sufficient	money	for

ltt	lhm̄	ʔt	kl	crkyhm̄
give.INF	to.them	ACC	all	needs.POSS.3PL.M

'There are parents who don't have enough money to **give** them all their needs'

It is important to acknowledge that, while annotators were expected to follow the faithfulness principle, this principle does not provide a prescription for annotation since, ultimately, the annotation process is based on intuitions. More specifically, annotators were guided to rely on their initial intuition, rather than actively seek alternative formulations. Thus, choosing between alternative formulations was only

[14] The verb להספיק is ambiguous. One of its uses does take a direct object, but its meaning ('to succeed in doing something on time') is incompatible with the given context.

relevant when the annotators thought of such alternatives spontaneously. As a result, annotators may occasionally apply corrections that are not optimal in terms of faithfulness, and there is no effective way to detect such cases. In Sect. 5.3, we attempt to evaluate the inter-annotator variability, part of which can be attributed to variability in the precision of applying the faithfulness principle.

5.1.5 Uncertainty

In many cases, the author expresses an idea in a way that is atypical of native language, and there is some uncertainty about the appropriate correction. In some of these cases, the intended meaning seems clear but there are several, equally plausible, alternative ways of expressing the idea in the target language. In such cases, annotators could specify multiple target hypotheses in their annotation. For example, sentence (6) is awkward, if not ungrammatical. Two equally plausible target hypotheses of (6) are given in (6′) and (6″). (6′) preserves the verb from the original sentence, while (6″) preserves the root of the object ע-נ-י/ה (ʕ-n-y/h) but embeds it in the verb rather than in the object.

(6) **נותן לו את העניין** אף אחד לא מסתכל על האחר או[???]

ʔp̄	ʔxd	lʔ	mstkl	ʕl	h-ʔxr	ʔw	**nwtn̄**	**lw**	**ʔt**
even	one	NEG	look.SG.M.PRS	on	the-other	or	**give.SG.M.PRS**	**to.him**	ACC

h-ʕnyyn̄
the-interest

[???] 'No one looks at the other or **gives the interest to them**'

(6′) **נותן לו את תשומת הלב** אף אחד לא מסתכל על האחר או

ʔp̄	ʔxd	lʔ	mstkl	ʕl	h-ʔxr	ʔw	**nwtn̄**	**lw**	**ʔt**
no	one	NEG	look.SG.M.PRS	on	the-other	or	**give.SG.M.PRS**	**to.him**	ACC

tšwmt **h-lb**
input.CONSTR **the-heart**

'No one looks at the other or **gives attention to them**'

(6″) **מתעניין בו** אף אחד לא מסתכל על האחר או

ʔp̄	ʔxd	lʔ	mstkl	ʕl	h-ʔxr	ʔw	**mtʕnyyn̄**	**bw**
no	one	NEG	look.SG.M.PRS	on	the-other	or	**take interest.SG.M.PRS**	**in.him**

'No one looks at the other or **takes interest in them**'

In HELEECS, multiple target hypotheses are indicated in separate columns, as shown in Table 6. The Token column corresponds to sentence (6). TH1 is a modified version of the full text (e.g., sentence 6′), while TH2 indicates only alternatives to corrections made in TH1 (e.g., parts of 6″ that are different from 6′), and is otherwise empty.

Another kind of uncertainty occurs when the intended meaning is unclear. In such cases, annotators were advised to leave the text uncorrected and, instead,

Table 6 Multiple target hypotheses

Token	TH1	TH2
ʔp̄	ʔp̄	
ʔxd	ʔxd	
lʔ	lʔ	
mstkl	mstkl	
ʕl	ʕl	
h-ʔxr	h-ʔxr	
ʔw	ʔw	
nwtn̄	**nwtn̄**	**mtʕnyyn̄**
lw	**lw**	**bw**
ʔt	**ʔt**	&&
h-ʕnyyn̄	**tšwmt**	&&
&&	**h-lb**	&&

make free-form comments, or assign special error tags to parts of the text during the error annotation process (see Sect. 5.2.4). For example, consider the following sentence:

(7) *נראה לי שיש משהו שדומה כמו **החלה** טכנולוגיה

nrʔh	ly	š-yš	mšhw	š-dwmh	kmw	**hxlh**	ṭknwlwgyh
seem.SG.M.PRS	to.me	that-exist	something	that-similar.SG.M	like	**applying**	technology

*'It seems to me that there is something that is similar like technology **applying**'

The phrase "החלה טכנולוגיה"<hxlh ṭknwlwgyh>'technology applying' is ungrammatical, but it is not clear what the intended meaning was. If the author meant 'application of technology' then some grammatical change is required: either insertion of the preposition של<šl>'of' between the words, or changing the conjugation of החלה to the construct state (i.e., **ת**החל<hxl**t**>'application.CONSTR'). In addition, it seems that some information is missing (e.g., 'application of technology' to what?). In fact, it is not clear at all that the author meant to use the word "החלה" 'application' (which is a rather high register word), but rather some other semantically, morphologically, or phonologically similar word. There is not enough information in the sentence to help recover the target word. The word דומה<dwmh>'similar.SG.M' suggests a comparison between entities, which could potentially be helpful. However, the compared entities are not mentioned in the sentence and, since the original order of the sentences is unknown, the context cannot help in determining what the relevant entities are. In this case, the most suitable solution would be to leave the text unaltered, and make comments about the problems in the sentence.

5.2 Interpretation

After revising a text (i.e., forming the target hypothesis), the deviations between the original and revised text were analyzed and tagged. The error tags are stored in a

separate column alongside the columns of original and revised tokens. If a single token contains multiple independent errors (e.g., a spelling error and a syntactic error), each error is tagged in a separate error column. If there are multiple target hypotheses for a given phrase, each one has its own set of error annotation columns.

Table 7 demonstrates revision and error annotation of a sentence. The Token column contains the tokenization of the original sentence (8), the TH1 column contains the tokenization of the revised sentence (8′), and the columns labelled Error1_TH1 and Error2_TH1 contain the error tags. The full list of error tags used in this project is included in an appendix supplied with the online corpus. Note that tilde signs in glosses indicate deliberate translation misspells (e.g., *teknology*) that mirror orthographic errors in the Hebrew text.

(8) כל דקה שלא מסתכל **את הטלפון** **משהו** למצוא יותר קשה **זה** יותר **הטחנולוגיה מתפתח***

ywtr	**h-ṭxnwlwgyh**	**mtptx**	ywtr	qšh	**zh**	lmcw?	**mšhw**
more	**~the-teknology**	**evolve.SG.M.PRS**	more	hard.SG.M	**this**	find.INF	**something**

š-l?	mstkl	**ʔt**	**h-ṭlpwn̄**	kl	dqh	
that-NEG	look.SG.M.PRS	ACC	**the-telephone**	every	minute	

*‘**More the teknology(F) evolves.M** more difficult it is to find **something** that doesn’t look **the phone** every minute’

(8′) כל דקה שלא מסתכל **בטלפון** **מישהו** יותר קשה למצוא **ככל שהטכנולוגיה מתפתחת**

kkl	**š-h-ṭknwlwgyh**	**mtptxt**	ywtr	qšh	lmcw?	**myšhw**
as much	**that-the-technology**	**evolve.SG.F.PRS**	more	hard.SG.M	find.INF	**someone**

š-l?	mstkl	**b-ṭlpwn̄**	kl	dqh	
that-NEG	look.SG.M.PRS	**in.the-telephone**	every	minute	

‘**The more the technology(F) evolves.F** the more difficult it is to find **someone** that doesn’t look **at the phone** every minute’

5.2.1 Basic error classification

Error tags have the general form of *function(arguments)*. This enables tagging a wide array of errors with a relatively small basic vocabulary of codes. In this configuration, *functions* indicate the nature of the deviation between the original and the revised token. Some common types of functions include: `miss` (a missing element), `redun` (a redundant element), and `wrong` (a wrong element). *Arguments* to the functions usually denote linguistic categories affected by the error. These categories include, among other things: orthographic elements, various categories of function words (e.g., prepositions, conjunctions), syntactic categories (e.g., subject, predicate), and categories of content words (e.g., noun, adjective). Most functions require only a single argument. For example, row 1 in Table 7 demonstrates tagging of an incorrect conjunction.

Other error functions require two arguments. This configuration is typically used with agreement errors. In these cases, the arguments to the function denote the

Table 7 Tokenized, revised and annotated text

	Token	TH1	Error1_TH1	Error2_TH1
1	ywtr	kkl	`wrong(conj)`	
2	h-ṭxnwlwgyh	š-h-ṭknwlwgyh	`shouldB(כ,ח)`	`miss(conj,##)`
3	mtptx	mtptxt	`agree(subj,pred)`	
4	ywtr	ywtr		
5	qšh	qšh		
6	zh	&&	`redun(dem)`	
7	lmcwʔ	lmcwʔ		
8	mšhw	myšhw	`oMiss(י)`	
9	š-lʔ	š-lʔ		
10	mstkl	mstkl		
11	ʔt	&&	`wrong(prep,&&)`	
12	h-ṭlpwn̄	b-ṭlpwn̄	`wrong(prep)`	
13	kl	kl		
14	dqh	dqh		

Tags legend: `wrong` = incorrect element, `conj` = conjunction, `shouldB` = element 1 should be element 2, `miss` = missing element, `agree` = agreement error, `subj` = subject of clause, `pred` = predicate, `redun` = redundant element, `dem` = demonstrative, `oMiss` = missing letter, `prep` = preposition

categories of the two elements for which there is a lack of agreement in gender, number, or person. For instance, row 3 in Table 7 contains the tag `agree(subj,pred)`, indicating an agreement error between the feminine subject of the clause, טכנולוגיה < ṭknwlwgyh > 'technology' and the main predicate of the clause מתפתח < mtptx > 'evolve.SG.M.PRS', which is masculine.

5.2.2 Multiple analyses

If there is more than one likely analysis of a given error, alternative analyses can be indicated side-by-side. For example, (9) contains the word form יוכלים < ywklym̄ >, which does not exist in Hebrew. In (9′), it was corrected to יכולים < ykwlym̄ > 'can.PL.M.PRS', resulting in a grammatical sentence.

(9) *הצעירים לא **יוכלים** לעבוד

h-cʕyrym̄	lʔ	**ywklym̄**	lʕbwd
the-young.PL.M	NEG	~**abel**	work.INF

(9′) הצעירים לא **יכולים** לעבוד

h-cʕyrym̄	lʔ	**ykwlym̄**	lʕbwd
the-young.PL.M	NEG	**able.PL.M.PRS**	work.INF

'The young are **unable** to work'

Table 8 Alternative error analyses

Token	TH1	Error1_TH1	TH2	Error1_TH2
h-cʕyrym̄	h-cʕyrym̄			
lʔ	lʔ			
ywklym̄	ykwlym̄	`metathesis(כו)`	ykwlym̄	`wrong(pattern)`
lʕbwd	lʕbwd			

The error in this example can be analyzed on two different levels: at the orthographic level it can be analyzed as metathesis of two adjacent letters (i.e., correct: כו → incorrect: וכ). Alternatively, it can be analyzed as a morphological error, i.e., selection of an incorrect non-linear pattern. Both the orthographic and morphological accounts are plausible. Table 8 demonstrates alternative analyses of the same error in HELEECS. The TH1 column contains the full revised text, as explained earlier. The TH2 column contains a copy of the revised token <ykwlym̄> (this is in contrast to situations described in Sect. 5.1.5, in which TH2 was different from TH1). Alternative analyses of the error are indicated in the Error1_TH1 and Error1_TH2 columns.

5.2.3 Dependent corrections

Occasionally, correction of one error entails additional corrections, often in different tokens (i.e., some corrections are dependent on others). While we tagged every correction made in the corpus, dependent corrections were excluded from statistical analysis in order to avoid overestimation of the number of errors in the corpus.

One type of dependent correction that was not counted involved insertion of a dummy token that accompanied additional modifications. As explained in Sect. 4.3, in cases such as deletion, insertion, splitting, merging, or movement of words, a dummy token && was inserted in order to maintain the alignment between original and revised tokens. However, this action may result in differences between the token columns on several rows (some reflecting true errors, others reflecting corrections of alignment). Since the multiple differences stem from a single error, counting all these rows would lead to an overestimation of the number of errors. To prevent this overestimation, we used the same error code in all the rows affected by the same error and added `&&` as an argument to the error function in all the rows containing the dummy token &&.

For example, row 8 in Table 9 demonstrates the annotation of a dummy token inserted as part of a preposition correction in sentence (8) (see also Table 7). The correction replaced the stand-alone preposition את <ʔt> (an accusative marker) by the cliticized preposition ב ‘in’. Overall, the single preposition correction resulted in the change of two tokens. The change in row 9 was tagged `wrong(prep)`, while

Table 9 Error tags and dummy tokens

	Token	TH1	Error1_TH1
1	ywtr	ywtr	
2	qšh	qšh	
3	zh	&&	`redun(dem)`
4	lmcwʔ	lmcwʔ	
5	mšhw	myšhw	`oMiss(י)`
6	š-lʔ	š-lʔ	
7	mstkl	mstkl	
8	ʔt	&&	`wrong(prep,&&)`
9	h-ṭlpwn̄̄	b-ṭlpwn̄̄	`wrong(prep)`
10	kl	kl	
11	dqh	dqh	

Note that not all dummy tokens were tagged with `&&`. Row 3 in Table 9 demonstrates deletion of the demonstrative <zh> 'this.M'. A dummy token was inserted in the TH1 column to maintain alignment between original and revised texts, but the error tag, `redun(dem)` does not contain `&&` since the error correction affected only a single row, which is equal to the actual number of errors

the change in row 8, which contains the dummy token, was tagged `wrong(prep,&&)`. Thus, every row containing different original and revised tokens was tagged, but multiple tags related to the same error were marked to be excluded from further analysis.

Another case of uncounted error tags are those marking changes that are required due to other obligatory changes. We call such changes "chain corrections". Chain corrections do not correct things that were considered errors in the original text, but rather things that would have been errors after the application of another correction. We view chain corrections as stemming from a single source and do not count them in order to avoid overestimation of the number of errors in the corpus. Chain corrections are marked in HELEECS by passing `##` as an argument to the relevant error function.

One type of chain correction is related to a repeated error in multiple linked words. One such common case is a consistent incorrect usage or omission of a grammatical element in coordination or list constructions. In such a case, `##` is added to all repeated instances of the tag referring to the relevant error. For example, in (10) an incorrect preposition ב <b-> 'in' is repeated instead of the preposition ל <l-> 'to' (as in 10′). Since the errors are identical and occur in a coordination construction that complements a single predicate, we consider them as a single error. Consequently, the second occurrence of the error is tagged `wrong(prep,##)` to indicate that it is dependent on the first occurrence (see Table 10).

Table 10 Annotation of a "chain correction" in a coordination construction

Token	TH1	Error1_TH1
w-šmym̄	w-šmym̄	
lb	lb	
ywtr	ywtr	
b-ʕbwdh	l-ʕbwdh	`wrong(prep)`
w-lʔ	w-lʔ	
b-mšpxh	l-mšpxh	`wrong(prep,##)`

(10) אנשים רוצים להצליח בחיים ושמים לב יותר **בעבודה** ולא **במשפחה***

ʔnšym̄	rwcym̄	lhclyx	b-xyym̄
people	want.PL.M.PRS	succeed.INF	in.the-life

w-šmym̄	lb	ywtr	**b-ʕbwdh**	w-lʔ	**b-mšpxh**
and-put.PL.M.PRS	heart	more	**in.the-work**	and-NEG	**in.the-family**

*'People want to succeed in life and pay more attention **in work** and not **in the family**'

(10′) אנשים רוצים להצליח בחיים ושמים לב יותר **לעבודה** ולא **למשפחה**

ʔnšym̄	rwcym̄	lhclyx	b-xyym̄
people	want.PL.M.PRS	succeed.INF	in.the-life

w-šmym̄	lb	ywtr	**l-ʕbwdh**	w-lʔ	**l-mšpxh**
and-put.PL.M.PRS	heart	more	**to.the-work**	and-NEG	**to.the-family**

'People want to succeed in life and pay more attention **to work** and not **to the family**'

Another type of chain correction involves reattachment of clitics. Recall that Hebrew has several function clitics. Occasionally, error correction requires such a clitic to be detached from one word and reattached to another. This results in changes in two words although there is only a single underlying error. These changes are tagged using complementary operators (i.e., `miss` and `redun`), and one of the tags includes `##` to indicate that the errors are dependent. For example, the phrase השאר דברים <h-šʔr dbrym̄> 'the rest of things' in (11) has the structure of a construct state (i.e., a noun modified by another noun). In definite construct states in formal Hebrew, the definite article should be attached to the modifier (i.e., the second noun) rather than to the modified (i.e., first) noun. When the definite article is attached to the modified noun, the appropriate correction requires the definite article to be detached from the first noun and reattached to the second, as in (11′). However, since these modifications are dependent, we tag the second correction with `##`, as in Table 11, to avoid inflating the number of estimated errors in the corpus.

Table 11 Annotation of clitic reattachment

Token	TH1	Error1_TH1
h-šʔr	šʔr	`redun(det)`
dbrym̄	h-dbrym̄	`miss(det,##)`

(11) *לטפל בכל **השאר דברים** בבית

lṭpl	b-kl	**h-šʔr**	**dbrym̄**	b-byt
handle.INF	in-all	**the-rest**	**things**	at.the-house

(11′) לטפל בכל **שאר הדברים** בבית

lṭpl	b-kl	**šʔr**	**h-dbrym̄**	b-byt
handle.INF	in-all	**rest**	**the-things**	at.the-house

‘To take care of **the rest of the stuff** at home’

5.2.4 Error tags and no correction

In many cases, a text clearly deviates from typical native language, but there is uncertainty about the appropriate correction. This can occur when the text is incomprehensible, or when there are several plausible corrections, each requiring a different major modification (e.g., change of syntactic structure). In such cases, annotators were advised not to correct the text. Yet, we tagged errors in individual words if the nature of the error was clear enough. We marked errors that did not accompany any revision of the text by adding `$$` as an argument to the error function.

An example of an uncorrected but tagged sentence can be seen in (12). The sentence is clearly incomplete. A possible correction would be to insert some modal adjective, such as עדיף <ʕdyp̄> ‘preferable’ as in (12′). However, it is unclear whether that was the author’s intention. Therefore, an alternative solution would be to insert a dummy token in both original and revised token columns and tag the error: `miss(lex,$$)`, i.e., a missing unknown lexical item. This is demonstrated in Table 12.

(12) *לדעתי להיכנס לנושא שיותר קל לך

ldʕty	lhykns	l-nwšʔ	š-ywtr	ql	lk̄
in.my.opinion	enter.INF	to-subject	that-more	easy.SG.M	to-you.SG

*‘In my opinion to get into a subject that is easier for you’

(12′) לדעתי **עדיף** להיכנס לנושא שיותר קל לך

ldʕty	**ʕdyp̄**	lhykns	l-nwšʔ	š-ywtr	ql	lk̄
in.my.opinion	**preferable**	enter.INF	to-subject	that-more	easy.SG.M	to-you.SG

‘In my opinion it is better to get into a subject that is easier for you’

5.2.5 Higher-level interpretations

Another feature of the annotation scheme used in this corpus is the inclusion of interpretive (or, explanatory) error tags. In many cases, an error on one level of

Table 12 Error tagging of an unknown missing lexical item

Token	TH1	Error1_TH1
ldʕty	ldʕty	
&&	&&	`miss(lex,$$)`
lhykns	lhykns	

analysis affects higher linguistic levels as well. In other cases, scrutinizing an error reveals a plausible cognitive cause for the error, which is not captured by the surface description of the error. In such cases, annotators were able to use an additional set of interpretive error tags to specify their observations. The interpretive error tags were added to the annotation separately (i.e., in distinct columns) from the other error tags.

One class of interpretive error tags analyzes the cognitive basis of orthographic errors. Most often, such errors are analyzed from a phonological perspective (i.e., influence of pronunciation on the written form) or from a visual perspective (i.e., substitution of similarly looking letters). Another class of interpretive error tags analyzes lexical and syntactic errors. The analysis can indicate details such as the semantic effect of a wrong lexical item (e.g., selection of a semantically-related, but inappropriate, word), pragmatic effects (inconsistent use of grammatical tense or person throughout a sentence), and even the use of inappropriate register.

For instance, sentence (13) demonstrates several errors that can be analyzed from different perspectives. The corrected sentence is shown in (13′). Table 13 displays the analysis of the errors, where columns labelled Error specify the more basic description of errors, and columns labelled Interp contain interpretations of individual errors relative to a specific target hypothesis (e.g., Interp2_TH1 is an interpretation of the second error analyzed in target hypothesis 1).

Table 13 Annotated text with explanatory error tags

	Token	TH1	Error1_TH1	Interp1_TH1	Error2_TH1	Interp2_TH1
1	bgll	bgll				
2	kkh	zh	`wrong(dem)`	`colloc`		
3	ʔny	ʔny				
4	yʕšh	ʔʕšh	`shouldB(א,י)`	`pronuncReg/register`		
5	bsykwmtry	psykwmṭry	`shouldB(פ,ב)`	`pronuncL2`	`shouldB(ט,ת)`	`homophone`

Tags legend: `wrong`=incorrect element, `dem`=demonstrative, `colloc`=miscollocation, `shouldB(x,y)`=element x should be element y, `pronuncReg`=regular pronunciation (of native speakers), `register`=inappropriate register, `pronuncL2`=pronunciation of L2 speakers (with a specific L1), `homophone`=homophonic letter substitution

(13) *בגלל **ככה** אני **יעשה בסיכומתרי**

bgll	**kkh**	ʔny	**yʕšh**	**bsykwmtry**
because of	**this way**	I	**do.3SG.M.FUT**	**~bsychomettric** (test)

*‘Because of **this way** I **will take (3SG.M)** the **bsychomettric** (test)’

(13′) בגלל **זה** אני **אעשה פסיכומטרי**

bgll	**zh**	ʔny	**ʔʕšh**	**psykwmṭry**
because of	**this/that**	I	**do.1SG.FUT**	**psychometric** (test)

‘Because of **that** I **will take** the **psychometric** (test)’

The use of ככה < kkh > ‘this way’ instead of זה < zh > ‘this’ is analyzed as a wrong demonstrative (row 2). In addition, it can be viewed as a miscollocation – deformation of the collocation בגלל זה < bgll zh > ‘because of that’ (an anonymous reviewer suggested that < bgll kkh > ‘because-of this-way’ is a direct calque of Arabic [ʕaʃaːn heːk] ‘because-of so’, which is the idiomatic way of saying ‘because of that’ in colloquial Arabic).

The use of יעשה < yʕšh > ‘do.3SG.FUT’ instead of אעשה < ʔʕšh > ‘do.1SG.FUT’ is a case of letter substitution, which reflects the colloquial pronunciation of the word, and is common even in the writing of native speakers (row 4). Moreover, it is noteworthy that even if the error is tolerable in informal writing, it is inappropriate in formal (e.g., essay) writing. Thus, the error can be further analyzed as a register error. Overall, the letter substitution in this example has two relevant aspects: a direct reflection of the common pronunciation of the word and a failure to use the appropriate register in formal writing. Both aspects can be indicated in the annotation by concatenating the relevant codes with a separating slash in the interpretation column (i.e., `pronuncReg/register`).

Finally, בסיכומתרי < bsykwmtry > ‘~bsychomettric (test)’ exhibits two spelling errors (row 5). The פ-ב substitution is a common error in the Hebrew of native speakers of Arabic resulting from the absence of the consonant [p] (represented by the letter פ) in Arabic (Abu Baker, 2016). Thus, it can be analyzed as an error reflecting the common pronunciation of L2 Hebrew speakers (with Arabic L1). The ט-ת substitution is a homophonic letter substitution (both letters represent the consonant [t]). This could also be influenced by the Arabic orthographic form سيكومتري [siːkuːmitrij] ‘psychometric’, where [t] is represented by the letter ت, which is the equivalent of Hebrew ת, not ט.[15]

In summary, the interpretive error tags represent a more speculative analysis, and can provide valuable insights that would be harder to reach without specific research hypotheses.

[15] We thank an anonymous reviewer for suggesting this analysis.

5.3 Evaluation

To evaluate the quality of the corrections and annotations, we chose 54 essays, at various proficiency levels and across all three L1s, to be annotated and corrected by two experienced annotators. In total, this evaluation set included 428 sentences comprising 7757 tokens. The size of the evaluation set (in terms of the number of essays, sentences, and tokens) is 5% of the size of the annotated corpus. The number of words corrected by both annotators was 671, about 9% of all tokens in the evaluation set.

Due to the complexity of the annotation process, the notion of inter-annotator agreement became complex as well. We calculated inter-annotator agreement on several levels: (i) whether annotators agreed that some word or expression contained an error, (ii) whether they applied the same correction, and (iii) whether they annotated the error similarly when the correction was identical. All cases of disagreement between annotators in these files were resolved by consultation with a third annotator.

The first inter-annotator agreement measure looked only at the binary question, whether both annotators treated word tokens in the same way (i.e., left untouched or corrected). The agreement between the two annotators (micro-averaged over all essays) was 95.4% (Range: 90%–99%, SD: 2%); the macro-average was 95.6%.

A second, harsher measure looked at the proportion of tokens that were corrected identically by both annotators. This measure takes into account (in other words, penalizes disagreement on) both the binary decision (whether to correct a token) and the actual correction. That is, the second agreement measure is the number of tokens corrected identically by both annotators divided by the number of tokens corrected by either annotator. Here, since the annotators had more freedom in determining the target hypothesis of an erroneous token, the agreement was only 57% (Range: 11%–83%, SD: 15%); the macro-average was 58%.

To understand why the agreement level on the corrections was relatively low, we scrutinized all cases of disagreement. Overall, we identified four types of disagreement. The distribution of correction differences over the various types is listed in Table 14.

Differences due to different target hypotheses are cases in which the annotators chose different but valid ways to correct the texts. Such differences reflect the natural variability of the language (see also Sect. 5.1.5 above). For example, the phrase in (14) is ungrammatical. Both annotators corrected the word הראשון<h-rʔšwn̄> 'the-first.SG.M', but each applied a different but equally acceptable correction (see 14′ and 14″).[16]

Table 14 Categories of disagreements on corrections

Type	%
Different target hypothesis	47
Annotator error	26
Differences in chain corrections	24
Partially overlapping corrections	3

[16] As an anonymous reviewer pointed out, it could be argued that (14′) is a preferable correction to (14″) according to the faithfulness principle, since (14′) involves only a morphological change, while (14″) involves both a morphological and a lexical change. However, it should be noted that the two sentences have slightly different meanings and both are plausible target hypotheses of (14).

(14) *מטרת הראשון של האפליקציות

mṭrt	**h-rʔšwn̄**	šl	h-ʔplyqcywt
goal(F).CONSTR	**the-first.SG.M**	of	the-applications

*‘**The goal(F) of first(M)** of the applications’

(14′) **המטרה הראשונה** של האפליקציות

h-mṭrh	**h-rʔšwnh**	šl	h-ʔplyqcywt
the-goal(F)	**the-first.SG.F**	of	the-applications

‘**The first goal** of the applications’

(14″) **המטרה המקורית** של האפליקציות

h-mṭrh	**h-mqwryt**	šl	h-ʔplyqcywt
the-goal(F)	**the-original.SG.F**	of	the-applications

‘**The original goal** of the applications’

The second type of disagreement was due to an error on part of one of the annotators. Most often, the error was failing to correct an obvious error in the text (e.g., a spelling error). Such errors cannot be prevented completely, but it is important to estimate their frequency and overall effect on the annotations.

The third type of disagreement was due to differences in chain corrections. As discussed in 5.2.3, chain corrections refer to a series of corrections in a multi-word phrase, such that a correction of one word requires corrections of additional words in the phrase. If the annotators disagreed on the first correction, this could lead to further disagreements. The disagreement on the first word is analyzed according to one of the previous categories (different target hypothesis, annotator error). However, the additional disagreements should be counted separately, since the words in the phrase are inter-dependent. A common case of disagreement in chain corrections involves the alternation between free and bound morphemes that are semantically equivalent. For example, (15) uses an inappropriate phrase to denote causality. Both annotators corrected it by adding a conjunction. However, the correction in (15″) also required an omission of a bound preposition מ<m-> ‘from’, resulting in a difference in two tokens between the two corrections, as demonstrated in Table 15.

Table 15 Disagreement in chain corrections

Token	Annotator1	Annotator2	
ʔnšym̄	ʔnšym̄	ʔnšym̄	
š-htʔbdw	š-htʔbdw	š-htʔbdw	
&&	k-tvcʔh	bgll	
m-ʔtry	m-ʔtry	ʔtry	←chain correction
ʔynṭrnṭ	ʔynṭrnṭ	ʔynṭrnṭ	

(15) *אנשים שהתאבדו **מאתרי** אינטרנט

ʔnšym̄	š-htʔbdw	**m-ʔtry**	ʔynṭrnṭ
people	that-commit suicide.PL.PST	**from-sites.CONSTR**	internet

*‘People who committed suicide **from** websites’

(15′) אנשים שהתאבדו **כתוצאה מאתרי** אינטרנט

ʔnšym̄	š-htʔbdw	**k-twcʔh**	**m-ʔtry**	ʔynṭrnṭ
people	that-commit suicide.PL.PST	**as-result**	**from-sites.CONSTR**	internet

‘People who committed suicide **as a result of** websites’

(15″) אנשים שהתאבדו **בגלל אתרי** אינטרנט

ʔnšym̄	š-htʔbdw	**bgll**	**ʔtry**	ʔynṭrnṭ
people	that-commit suicide.PL.PST	**because of**	**sites.CONSTR**	internet

‘People who committed suicide **because of** websites’

The last type of disagreement on corrections was in partially overlapping corrections. This type refers to cases of multiple errors in a single word where the annotators agreed on the correction of some of the errors, but not on the others. For example, both annotators changed the bound preposition ל<l-> ‘to’ in (16) to the bound preposition ב<b-> ‘in’. However, the first annotator did not make additional changes (16′), while the second annotator also changed the noun to which the bound preposition is attached (16″). Thus, the annotators disagreed at the token level (<b-dbr> ‘in-thing’ vs. <b-mšhw> ‘in-something’). However, the fact that they did agree on the correction of the preposition should not be overlooked. The annotations for the token in question in (16′) and (16″) are compared in Table 16. Both annotators (marked “An1” and “An2”, respectively) used the tag `wrong(prep)` to mark the incorrect preposition, but annotator 2 also used the tag `wrong(lex)` to mark the choice of noun.

(16) *אם אדם חפץ בכל ליבו **לדבר** הוא יגשים אותו

ʔm̄	ʔdm̄	xpc̄	b-kl	lybw	**l-dbr**	hwʔ	ygšym̄	ʔwtw
if	person	wish.SG.M.PRS	in-all	heart.POSS.3SG.M	**to-thing**	he	fulfill.3SG.M.FUT	ACC.3SG.M

*‘If someone wishes with all their heart **to a thing** they will achieve it’

(16′) אם אדם חפץ בכל ליבו **בדבר** הוא יגשים אותו

ʔm̄	ʔdm̄	xpc̄	b-kl	lybw	**b-dbr**	hwʔ	ygšym̄	ʔwtw
if	person	wish.SG.M.PRS	in-all	heart.POSS.3SG.M	**in-thing**	he	fulfill.3SG.M.FUT	ACC.3SG.M

‘If someone wishes with all their heart **for a thing** they will achieve it’

(16″) אם אדם חפץ בכל ליבו **במשהו** הוא יגשים אותו

ʔm̄	ʔdm̄	xpc̄	b-kl	lybw	**b-mšhw**	hwʔ	ygšym̄
if	person	wish.SG.M.PRS	in-all	heart.POSS.3SG.M	**in-something**	he	fulfill.3SG.M.FUT

ʔwtw

ACC.3SG.M

‘If someone wishes with all their heart **for something** they will achieve it’

Table 16 Partially overlapping corrections

Token	TH1_An1	TH1_An2	Error1_An1	Error1_An2	Error2_An2
l-dbr	b-dbr	b-mšhw	`wrong(prep)`	`wrong(prep)`	`wrong(lex)`

To summarize, when analyzing learner texts that have been corrected, one should keep in mind that the corrections do not represent an absolute truth. First, corrected texts may still contain errors. This includes grammatical errors that would be considered errors by any standard, but also expressions that could be considered errors in some register or dialect but not in another. Second, a given correction could be only one of several plausible corrections that was chosen by a specific annotator. The annotation guidelines attempt to minimize such inconsistency (e.g., by including alternative corrections in the annotated text), but some variability in the corrections cannot be avoided. For example, it could be argued that (16′) is a preferable correction to (16″) according to the faithfulness principle, since (16′) changes only a preposition, while (16″) also makes a lexical change.[17] Yet, note that both formulations are grammatical and similar in meaning. Thus, preferring one solution to the other is to some extent a matter of an arbitrary decision. Overall, this example demonstrates that some inter-annotator variability cannot be avoided. At best, one can attempt to assess the amount of inter-annotator variability and propose additional guidelines that might reduce it.

Next, we discuss the third inter-annotator agreement measure, which was calculated based on the annotations of tokens that were corrected identically by the annotators. Instead of using the actual error tags, we used more general classes of tags, e.g., one class that accounts for all errors involving prepositions (missing, redundant, and wrong prepositions). The overall agreement on the annotations was 80% (Range 0–100%, SD: 18%). As in the analysis of the corrections, we distinguished several types of disagreements on annotations. The distribution of differences in errors tags over the various types is listed in Table 17.

Differences in the interpretation of errors occur when there is more than one plausible way to analyze a given error. One of the most common cases of this type of disagreement involves errors in letters that represent function clitics, such as

Table 17 Categories of disagreements on annotations (percentage out of 99 tokens)

Type	%
Different interpretations	35
Annotator error	29
Annotation difference with no corrections	20
Partially overlapping annotations	16

[17] We thank an anonymous reviewer for suggesting this argument.

prepositions (see Sect. 3.2). Such errors could, in principle, be analyzed as orthographic errors or as syntactic errors. For example, in one case, both annotators corrected the word קרוב <qrwb> 'close' to בקרוב <bqrwb> 'soon' (lit. 'in close'). One of them analyzed the error in the original token as a missing letter, while the other analyzed it as a missing preposition.

As in the disagreements on the corrections, many of the annotation differences were due to an error on part of one of the annotators. Most often, this happened when both annotators applied the same correction, but one of them did not assign an error tag to the revised token.

The third type of disagreement includes cases in which neither of the annotators corrected a given word, but one of them assigned an error tag to it. This usually happened when a content word was used inappropriately, but there was no clear target hypothesis. In such cases, the annotation scheme enables annotators to tag a word even if it was left uncorrected (see Sect. 5.2.4). However, adding an error tag is optional in these cases, thus, one annotator may choose to tag the error, while the other may choose not to tag it.

The last type of disagreement on annotations is in partially overlapping annotations. These cases involve multiple errors in a single word where the annotators agreed on the annotation of some of the errors, but not on the others. One such case involves an error that both annotators corrected and annotated similarly and an additional error that neither annotator corrected, but one of them tagged nonetheless. For example, in one instance, both annotators corrected the word אש <ʔš> 'fire' to איש <ʔyš> 'man' and analyzed the error as a missing letter. However, one of them commented that even the corrected word was inappropriate in the context, and added an error tag for a wrong lexical item. The other annotator did not tag the lexical error leading to partial disagreement on the annotation. Since the annotators did agree on one of the errors, the inter-annotator agreement analysis should take this into account.

To conclude, learner language inevitably involves a certain degree of variability. In addition, there is some uncertainty in native speaker interpretation of learner language. Our annotation scheme and guidelines were designed with these facts in mind. On the one hand, we attempted to minimize the variability of the annotations by providing elaborate guidelines that address common issues encountered during the annotation process (see appendix). On the other hand, we acknowledged the fact that the variability cannot be eliminated completely. Consequently, we decided to incorporate the variability in the annotation architecture by allowing annotators to specify multiple target hypotheses and multiple error tags whenever there was more than one way to correct and analyze a given error.

6 Analysis

To demonstrate the utility of the corpus, we compared the number of errors in various linguistic categories across the three L1s (see Table 18). Each error tag was classified as belonging to one of the linguistic categories and we calculated

Table 18 Error tags across the three L1s

Category	Arabic	French	Russian	H(2)	Sig
Orthography					
General	2241	2944	3135	18.90	*** a-r, a-f
Morphology					
Tokenization	126	106	147	3.44	
Linear morphology: stem-affix	284	256	210	4.55	
Non-linear morphology: patterns	978	691	518	77.61	*** a-r, a-f, f-r
Syntax					
Agreement	760	731	705	1.39	
Argument structure	182	129	118	14.76	** a-r, a-f
Conjunctions	753	485	372	99.24	*** a-r, a-f, f-r
Construct state	43	49	26	5.39	
Copulas	147	70	91	27.68	*** a-r, a-f
Pronouns and demonstratives	159	133	96	11.50	** a-r
Determiners	322	389	925	158.11	*** a-r, f-r
Existentials	52	33	25	7.50	* a-r
Negation	22	21	20	0.19	
Order	182	124	133	4.24	
Prepositions	943	1066	667	60.39	*** a-r, f-r
Punctuation	82	72	55	0.76	
Questions	23	21	36	3.43	
Relative clauses	46	64	64	2.11	
Semantics and lexicon					
General	837	703	594	25.22	*** a-r, a-f
Tokens	50,304	48,893	47,213		
Errors	8232	7907	7495		

Notes: H(2) = the Kruskal–Wallis test statistic (df = 2)

* $p < 0.05$

** $p < 0.01$

*** $p < 0.001$

the total number of errors tagged for each category in every essay. We then used an Independent-Samples Kruskal–Wallis Test to compare the distributions of the number of error tags in every category across the L1s (H_0 = the distribution of each error type is equal across the three L1s). It should be emphasized that the main purpose of this article is to introduce a language resource rather than perform an in-depth analysis of learner language. Nevertheless, we report below some findings that demonstrate the utility of the resource, in the hope that other scholars use it for these kinds of analyses.

Table 18 compares the number of errors in each category across the three L1s. The numbers were normalized per 50,000 tokens in order to make them comparable (the actual numbers of tokens in all the essays from each L1 are shown at the bottom of the table). The Sig column marks by stars comparisons that were

statistically significant. In addition, it specifies significant pairwise comparisons (a – Arabic, f – French, r – Russian; e.g., a-f = a significant difference between the distributions of a certain error type in essays authored by L1 Arabic and L1 French authors). Finally, the table specifies the total number of errors tagged for each L1.

The analysis reveals significant differences in error patterns in several categories across the L1s. The most notable difference was with respect to determiners, where L1 Russian authors made significantly more errors than L1 Arabic and L1 French authors. This can be attributed to the fact that Russian lacks definite articles, unlike Arabic, French, and Hebrew. Thus, we were able to demonstrate that the error annotation process can reflect differences in linguistic properties across different languages.

Other notable differences across the L1s were found with respect to the use of conjunctions and non-linear morphological patterns, where L1 Arabic authors made more errors than L1 French authors who, in turn, made more errors than L1 Russian authors. In addition, L1 Russian authors made significantly fewer errors in the use of prepositions than both L1 Arabic and L1 French authors (who did not differ from each other). Finally, L1 Arabic authors made more lexical errors and fewer orthographic errors than both L1 French and L1 Russian authors.

The fact that Arabic speakers made fewer orthographic and more morphological errors than the other two author groups may be attributed to the similarities between Arabic and Hebrew. First, Hebrew and Arabic have many shared roots and similar morpho-orthographic systems. Recognizing Hebrew words that are cognates of Arabic words may help Arabic speakers learn the correct spelling of Hebrew words, especially when homophonic letters are concerned. Second, the Arabic and Hebrew morphological systems are similar but not identical. Thus, Arabic speakers need to suppress their morphological knowledge of Arabic when, e.g., conjugating verbs in Hebrew. Failing to do so leads to an excess of morphological errors, which can serve as evidence for interference from L1 Arabic on using L2 Hebrew. It can be hypothesized that this is the case with errors as in utterance (17). The verb is inappropriate in this context, and it's noteworthy that it is conjugated in the Hitpa'el verb pattern that includes a templatic ת < t > (i.e., להת**ת**חשב < lh**t**XŠB > 'consider'). The appropriate verb לחשוב < lXŠwB > 'think' is conjugated in the Pa'al verb pattern that doesn't include a templatic ת, as in (17′). Interestingly, in Arabic, the meaning of deep thinking is expressed by the verb نتفكر < n**t**FKR > that is conjugated in a pattern that includes a templatic ت < t >. Thus, the presence of a templatic t in the corresponding Arabic verb may explain the choice of the inappropriate Hebrew verb.

(17) *עלינו **להתחשב מאוד** לפני שאנחנו מחליטים

ʕlynw	**lhtxšb**	**mʔwd**	lpny	š-ʔnxnw	mxlytym̄
on.us	**consider.INF**	**very**	before	that-we	decide.PL.M.PRS

*'We need to **consider very** before we decide'

(17′) עלינו **לחשוב הרבה** לפני שאנחנו מחליטים

ʕlynw	**lxšb**	**hrbh**	lpny	š-ʔnxnw	mxlytym̄
on.us	**think.INF**	**a lot**	before	that-we	decide.PL.M.PRS

'We need to **think a lot** before we decide'

The different distributions of errors by L1 Arabic authors compared to the other groups can explain the annotators' impressions that essays by L1 Arabic authors tended to be harder to annotate and were more time-consuming (regardless of their grade). An excess of lexical, conjunction and morphological errors, as found in these texts, is likely to be detrimental to comprehension in terms of both content and logic. These effects are likely amplified by the sparse use of punctuation by L1 Arabic authors, which resulted in very long sentences. This can be concluded from the fact that the average number of tokens per essay was similar across the three L1s, while the average number of sentences per essay was considerably lower in essays authored by L1 Arabic speakers (see Table 3).

Interestingly, the finding regarding sentence length is not unique to our corpus. Table 19 compares the mean number of tokens per sentence in several corpora of native Arabic, French, and Russian from the Universal Dependencies project (Marneffe et al., 2021). It is evident that sentence length in the Arabic corpus is considerably higher than in the other two languages. Thus, we may hypothesize that the higher sentence length in essays by L1 Arabic speakers in our corpus reflects the style of writing in their L1.[18]

Compared to the error patterns in essays by L1 Arabic authors, the major error types found in essays by L1 French and Russian authors seem to be less detrimental to comprehension. Orthographic errors alone are not expected to affect comprehension much, assuming the target word is appropriate. Determiner errors tend to make sentences sound "accented" but not incomprehensible. Preposition errors are also not expected to reduce comprehension much, since non-spatial/temporal

Table 19 Sentence length in Arabic, French, and Russian in several Universal Dependencies corpora

	Arabic	French	Russian
Sentences	28,402	29,735	111,238
Tokens	892,098	619,012	1,830,033
Tokens per sentence	31.41	20.82	16.45

Note: The corpora included in the analysis are:

Arabic: UD_Arabic-NYUAD, UD_Arabic-PADT, UD_Arabic-PUD

French: UD_French-FQB, UD_French-GSD, UD_French-PUD, UD_French-ParTUT, UD_French-ParisStories, UD_French-Rhapsodie, UD_French-Sequoia

Russian: UD_Russian-GSD, UD_Russian-PUD, UD_Russian-SynTagRus, UD_Russian-Taiga

[18] We thank an anonymous reviewer of this paper for suggesting the analysis of sentence length in the UD corpora.

prepositions are usually arbitrary and add little information beyond what is contained in content words.

Another type of analysis was performed by Nguyen and Wintner (2022), who conducted some basic classification experiments with the corpus. They were able to demonstrate that simple, feature-based classifiers can accurately distinguish between the native and the non-native authors; predict the native language of non-native writers; and quite accurately predict the non-natives' Hebrew proficiency scores, such that the model predictions were often indistinguishable from those of human raters. These results support the notion that there are strong, identifiable signals of the L1 and the authors' proficiency level in the corpus. We therefore trust that HELEECS will be invaluable both for research in learner language, including, for example, transfer effects from L1, and for practical educational applications.

7 Conclusions

We presented the Hebrew Learner Essay Corpus (HELEECS), a dataset of essays authored by native and non-native speakers of Hebrew. The dataset was computationally processed, is uniformly represented, and underwent error annotation. We expect it to be a valuable resource for any investigation of Hebrew as a second language, specifically when transfer effects from Arabic, French, and Russian are concerned. The corpus, the annotation scheme and the guidelines to the annotators are all available for research proposes.

At this time, only a third of the non-native essays in the corpus have been annotated. Further development of the corpus would include annotation of the remaining essays, as well as morpho-syntactic parsing and part-of-speech tagging of the non-native sub-corpus, and annotation, parsing and tagging of the native sub-corpus. Finally, additional analysis of the corpus can provide more insights regarding the influence of each of the L1s on performance in L2 Hebrew. The results of such analyses, combined with similar analyses of data from other languages, can shed light on universal and on language specific aspects of multilingualism.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10579-023-09712-w.

Acknowledgements We are immensely grateful to the Israeli National Institute for Testing and Evaluation for making the essays available. We are extremely grateful to Noam Ordan, Anke Lüdeling, Sarah Schneider, Isabelle Nguyen, and Dominique Bobeck for advice and fruitful discussions. We are also grateful to the anonymous reviewers for their constructive suggestions. We would like to thank Gur Meir for his help with the error tagging process. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 398186468 and by the Data Science Research Center at the University of Haifa.

Author contributions A.P. and S.W. supervised the study and are responsible for its conceptualization and funding acquisition. A.B-S-T coordinated the collection of the corpus and supervised its digitization and pre-processing. C.G., L.H-S., and H.K. developed the methodology (i.e., the annotation scheme). C.G. wrote the manuscript and performed the statistical analysis. L.H-S. and H.K. performed the data annotation. All authors reviewed and commented on the manuscript.

Funding Open access funding provided by University of Haifa.

Data availability All the data used in this research are available upon request. See footnote 10 and the accompanying datasheet.

Declarations

Conflict of interest The authors declare no competing interests. The authors have no relevant financial or non-financial interests to disclose.

References

Abu Baker, R. (2016). Hashpa'at leshon ha-em ha-'aravit al diburam ve-al ktivatam shel studentim arvim be-mixlala dovert aravit. *Ivrit be-Kavana T'hila* (pp. 63–69).

Ben-Dror, I., Frost, R., & Bentin, S. (1995). Orthographic representation and phonemic segmentation in skilled readers: A cross-language comparison. *Psychological Science*. https://doi.org/10.1111/j.1467-9280.1995.tb00328.x

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6*, 587–604. https://doi.org/10.1162/tacl_a_00041

Bentin, S., & Frost, R. (1987). Processing lexical ambiguity and visual word recognition in a deep orthography. *Memory & Cognition, 15*(1), 13–23. https://doi.org/10.3758/BF03197708

Bergsma, S., Post, M., & Yarowsky, D. (2012). Stylometric analysis of scientific articles. *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, proceedings of the conference* (pp. 327–337).

Berzak, Y., Reichart, R., & Katz, B. (2015). Contrastive analysis with predictive power: Typology driven estimation of grammatical error distributions in ESL. *Proceedings of the 19th conference on computational natural language learning* (pp. 94–102). https://doi.org/10.18653/v1/k15-1010

Bykh, S., & Meurers, D. (2012). Native language identification using recurring N-grams – Investigating abstraction and domain dependence. *Proceedings of COLING, 2012*, 425–440.

Dąbrowska, E. (2018). Experience, aptitude and individual differences in native language ultimate attainment. *Cognition, 178*, 222–235. https://doi.org/10.1016/j.cognition.2018.05.018

De Knop, S., & Meunier, F. (2015). The "learner corpus research, cognitive linguistics and second language acquisition" nexus: A SWOT analysis. *Corpus Linguistics and Linguistic Theory, 11*(1), 1–18. https://doi.org/10.1515/cllt-2014-0004

de Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational Linguistics, 47*(2), 255–308. https://doi.org/10.1162/COLI_a_00402

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching, 47*(2), 157–177. https://doi.org/10.1515/iral.2009.007

Estival, D., Gaustad, T., Pham, S. B., Radford, W., & Hutchinson, B. (2007). Author profiling for English emails. *Proceedings of the 10th conference of the Pacific Association for computational linguistics* (pp. 263–272).

Fabri, R., Gasser, M., Habash, N., Kiraz, G., & Wintner, S. (2014). Linguistic introduction: The orthography, morphology and syntax of semitic languages. In I. Zitouni (Ed.), *Natural language processing of semitic languages* (pp. 3–41). Springer. https://doi.org/10.1007/978-3-642-45358-8_1

Frost, R. (2012). Towards a universal model of reading. *Behavioral and Brain Sciences, 35*(05), 263–279. https://doi.org/10.1017/S0140525X11001841

Frost, R., Deutsch, A., Gilboa, O., Tannenbaum, M., & Marslen-Wilson, W. D. (2000). Morphological priming: Dissociation of phonological, semantic, and morphological factors. *Memory & Cognition, 28*(8), 1277–1288. https://doi.org/10.3758/BF03211828

Gadish, R. (2012). Transcription vs. transliteration. *Ha'Ivrit: A Journal for the Hebrew Language*, *60*(1–2), 43–60 (Hebrew).

Gafni, C. (2015). Child Phonology Analyzer: Processing and analyzing transcribed speech. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th international congress of phonetic sciences.* (pp. 1–5, paper number 531). ISBN 978-0-85261-941-4

Gafni, C., Prior, A., & Wintner, S. (2022). The Hebrew Essay Corpus. *Proceedings of the 13th conference on language resources and evaluation* (pp. 5580–5586).

Gafni, C., Yablonski, M., & Ben-Shachar, M. (2019). Morphological sensitivity generalizes across modalities. *The Mental Lexicon, 14*(1), 37–67. https://doi.org/10.1075/ml.18020.gaf

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., & Crawford, K. (2020). Datasheets for datasets. In *arXiv preprint* arXiv:1803.09010. http://arxiv.org/abs/1803.09010

Gilquin, G. (2008). Combining contrastive and interlanguage analysis to apprehend transfer: detection, explanation, evaluation. In G. Gilquin, S. Papp, & M. Belén Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. 3–33). Rodopi. https://doi.org/10.1163/9789001206204_002

Gilquin, G., & Paquot, M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction, 1*(1), 41–61.

Goldin, G., Rabinovich, E., & Wintner, S. (2018). Native language identification with user generated content. *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3591–3601). https://doi.org/10.18653/v1/d18-1395

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition, and foreign language teaching* (pp. 3–33). John Benjamins.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research, 1*(1), 7–24. https://doi.org/10.1075/ijlcr.1.1.01gra

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast: Papers from a symposium on text-based cross-linguistic studies*. Lund University Press.

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge Handbook of Learner Corpus Research*. Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.001

Grice, P. (1989). *Studies in the way of words*. Harvard University Press.

Gries, S. T. (2008). Corpus-based methods in analyses of second language acquisition data. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 406–431). Routledge.

Gries, S. T. (2015). Statistics for learner corpus research. In *The Cambridge Handbook of Learner Corpus Research* (pp. 159–181). Cambridge University Press.

Gries, S. T., & Deshors, S. C. (2015). EFL and/vs. ESL? A multi-level regression modeling perspective on bridging the paradigm gap. *International Journal of Learner Corpus Research, 1*(1), 130–159. https://doi.org/10.1075/ijlcr.1.1.05gri

Hermet, M., & Désilets, A. (2009). Using first and second language models to correct preposition errors in second language authoring. *Proceedings of the NAACL HLT workshop on innovative use of NLP for building educational applications* (pp. 64–72). https://doi.org/10.3115/1609843.1609853

Hirschmann, H., Lüdeling, A., Rehbein, I., Reznicek, M., & Zeldes, A. (2013). Underuse of syntactic categories in Falko. A case study on modification. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead* (pp. 223–234). Presses Universitaires de Louvain.

Jacobs, K., Itai, A., & Wintner, S. (2020). Acronyms: Identification, expansion and disambiguation. *Annals of Mathematics and Artificial Intelligence, 88*(5–6), 517–532. https://doi.org/10.1007/s10472-018-9608-8

Koppel, M., Schler, J., & Zigdon, K. (2005). Determining an author's native language by mining a text for errors. *Proceedings of the Eleventh ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 624–628). https://doi.org/10.1145/1081870.1081947

More, A., Seker, A., Basmova, V., & Tsarfaty, R. (2019). Joint transition-based models for morpho-syntactic parsing: Parsing strategies for MRLs and a case study from modern Hebrew. *Transactions of the Association for Computational Linguistics, 7*(2001), 33–48. https://doi.org/10.1162/tacl_a_00253

National Institute for Testing & Evaluation. (1986). *Hebrew Proficiency Test (YAEL)*. https://www.nite.org.il/other-tests/yael/?lang=en

National Institute for Testing & Evaluation. (2012). *Psychometric Entrance Test (PET)*. https://www.nite.org.il/psychometric-entrance-test/?lang=en

Nguyen, I., & Wintner, S. (2022). Predicting the proficiency level of nonnative Hebrew authors. *Proceedings of the language resources and evaluation conference* (pp. 5356–5365). https://aclanthology.org/2022.lrec-1.573

Norman, T., Degani, T., & Peleg, O. (2016). Transfer of L1 visual word recognition strategies during early stages of L2 learning: Evidence from Hebrew learners whose first language is either Semitic or Indo-European. *Second Language Research, 32*(1), 109–122. https://doi.org/10.1177/0267658315608913

Ornan, U. (2017). Perfect Latin conversion for Hebrew. *Lěšonénu: A Journal for the Study of the Hebrew Language and Cognate Subjects, 79*(1), 184–197. (Hebrew).

Prior, A., & Markus, E. (2014). Morphological activation in sentence context: When the root prevails over the meaning. *Language, Cognition and Neuroscience, 29*(9), 1180–1188. https://doi.org/10.1080/23273798.2014.920511

Ravid, D. (2020). Derivation. In R. A. Berman (Ed.), *Usage-based studies in modern Hebrew: Background, morpho-lexicon, and syntax* (pp. 203–264). John Benjamins.

Ravid, D., & Malenky, A. (2001). Awareness of linear and nonlinear morphology in Hebrew: A developmental study. *First Language, 21*, 25–56. https://doi.org/10.1177/014272370102106102

Reznicek, M., Lüdeling, A., & Hirschmann, H. (2013). Competing target hypotheses in the Falko Corpus: A flexible multi-layer corpus architecture. In *Automatic Treatment and Analysis of Learner Corpus Data* (pp. 101–123).

Sabourin, L., Stowe, L. A., & De Haan, G. J. (2006). Transfer effects in learning a second language grammatical gender system. *Second Language Research, 22*(1), 1–29. https://doi.org/10.1191/0267658306sr259oa

Schütze, C. T. (2016). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Language Science Press. https://doi.org/10.26530/oapen_603356

Schwarzwald, O. (2002). *Studies in Hebrew Morphology*. The Open University. (in Hebrew).

Share, D. L., & Bar-On, A. (2018). Learning to read a semitic Abjad: The triplex model of Hebrew reading development. *Journal of Learning Disabilities, 51*(5), 444–453. https://doi.org/10.1177/0022219417718198

Shimron, J. (2003). *Language processing and acquisition in languages of semitic, root-based, morphology*. John Benjamins.

Sprouse, J. (2009). Revisiting satiation: Evidence for an equalization response strategy. *Linguistic Inquiry, 40*(2), 329–341. https://doi.org/10.1162/ling.2009.40.2.329

Tetreault, J., Blanchard, D., & Cahill, A. (2013). A report on the first native language identification shared task. Aclweb.Org (pp. 48–57). http://www.aclweb.org/anthology/W13-1706

Tomokiyo, L. M., & Jones, R. (2001). You're not from 'round here, are you? Naive Bayes detection of non-native utterances. *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 1–8).

Tono, Y. (2003). Learner corpora: Design , development and applications. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the 2003 Corpus Linguistics conference* (pp. 800–809). University Centre for Computer Corpus Research on Language.

Tsvetkov, Y., Twitto, N., Schneider, N., Ordan, N., Faruqui, M., Chahuneau, V., Wintner, S., & Dyer, C. (2013). Identifying the L1 of non-native writers: The CMU-Haifa system. *Proceedings of the eighth workshop on innovative use of NLP for Building Educational Applications* (pp. 279–287). https://doi.org/10.1001/archophthalmol.2010.205

Vyatkina, N., Hirschmann, H., & Golcher, F. (2015). Syntactic modification at early stages of L2 German writing development: A longitudinal learner corpus study. *Journal of Second Language Writing, 29*, 28–50. https://doi.org/10.1016/j.jslw.2015.06.006

Zeldes, A., Howell, N., Ordan, N., & Moshe, Y. B. (2022). A second wave of UD Hebrew Treebanking and Cross-Domain Parsing. *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 4331–4344).