# Analyzing learner language: the case of the Hebrew essay corpus

**Chen Gafni** ( chen.gafni@gmail.com )
University of Haifa
**Livnat Herzig Sheinfux**
University of Haifa
**Hadar Klunover**
University of Haifa
**Anat Prior**
University of Haifa
**Shuly Wintner**
University of Haifa

**Research Article**

**Keywords:** Learner corpora, Hebrew, non-native language, Crosslinguistic influence, Educational applications

**Posted Date:** January 16th, 2023

**DOI:** https://doi.org/10.21203/rs.3.rs-2433887/v1

# Abstract

We present the Hebrew Essay Corpus: an annotated corpus of Hebrew language argumentative essays authored by prospective higher-education students. The corpus includes both essays by native speakers, written as part of the psychometric exam that is used to assess their future success in academic studies; and essays authored by non-native speakers, with three different native languages, 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 whose main goal is to make the texts amenable to automatic processing (morphological and syntactic analysis). 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 in general, and propose various ways for dealing with these challenges.


# 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, 2015, 1996). 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 paper, we present the Hebrew Essay Corpus:[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, 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) 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 paper thus makes two main contributions. The more specific one is the introduction of the *Hebrew Essay Corpus*. 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 standardize the analysis as much as possible. In addition, it includes general guidelines intended to increase the awareness of annotators to the issue of linguistic variability.

The structure of this paper is as follows: after reviewing some pertinent morphological and orthographic features of Hebrew in Section 2, we describe the corpus (Section 3) and the process of error correction and annotation (Section 4).

Section 5 describes two use cases of the corpus. We conclude in Section 6 with suggestions for future research.

[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 & Wintner (2022).

## 2. 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).

# 2.1 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 *abjadsystem*). Vowels are represented mainly by diacritical marks called *nikkud* placed above, below, or inside letters. 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). 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).

This phonetic ambiguity is partially resolved by the inclusion of vowel letters (Hebrew: [imot kri'a]; Latin: *matres lectionis*, lit. 'mothers of reading'), a set of four letters { , , , } that can represent both vowels and consonants. The use of vowel letters is subject to various constraints, usually morphological. Constraints affect both the phonetic values of the letters and whether they can (or should) be omitted in unpointed Hebrew. For example, the endonym of the Hebrew language [ivrit] contains two [i] vowels. In the standard written version, , the second [i] is represented by the vowel letter Yod , while the first vowel is not represented in the script. An alternative spelling that explicitly represents the first [i] (i.e., ), is considered sub-standard, although it is phonetically equivalent to the standard spelling.

Another phonetic ambiguity in the Hebrew orthography is due to the existence of several letters representing the same sound (homophonic letters), and letters with more than one phonetic value. For instance, the voiceless velar stop /k/ is represented by two letters: Kaf ( ) and Qof ( ). As a result, Modern Hebrew has pairs of homophonic words such as 'yes' and 'nest', both pronounced [ken]. In addition, the letter Kaf represents not only the consonant /k/ but also /x/. This duality of Kaf is due to a context-dependent phonological process of spirantization, by which the plosive consonant /k/ changes into the fricative consonant [x] after a vowel. Spirantization is most conspicuous in verbal paradigms. For instance, represents [k] in [liškav] 'lie down.inf', but [x] in [šaxav] 'lie down.3sg.m.pst'. To complicate things even further, the process of spirantization is not productive in Modern Hebrew, and many words that used to exhibit the k~x alternation do not show this alternation consistently nowadays. Thus, Hebrew learners face multiple challenges in learning the form-sound mapping of the Hebrew orthography (e.g., learning when [k] is represented by and when by , as well as learning when is pronounced as [k] and when as [x]).

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 ( , , , , ) vs. word non-final ( , , , , ) positions. Second, some pairs of Hebrew letters have similar visual appearance. For instance, the letters Heh (/h/) and Heth (/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 ( , , ) only differ in the length of their vertical lines. Such similarities across letters and context-dependent variation of letter form may be confusing for learners.

## 2.2 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 and represents an abstract concept that carries the primary meaning component of the word. 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 properties such as lexical category, tense, gender, aspect, and so on. For example, the consonantal root . . (/X.S.N/) stands for the concept of immunity, or strength.[3] It can be incorporated in various patterns to create words such as ([XiSuN], noun: 'vaccine'), ([hitXaSNu], verb: 'get vaccinated.3pl.pst') and ([XaSiN], adjective: 'immune.sg.m, resistant.sg.m').

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 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 /CiCeC/ and /CaCiC/.[4] The only visual difference between their written forms is the position of the letter representing the vowel /i/ (the vowels /e/ and /a/ are not represented by letters in the written form). /CiCeC/ is a verbal pattern, while /CaCiC/ is either adjectival or nominal. Importantly, 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 . . /M.H.R/, the obtained words are semantically-related: a verb (CiCeC: [MiHeR] 'rush.3sg.m.pst') and an adjective (CaCiC: [MaHiR] 'fast.sg.m'). In the case of the root . . /X.Z.R/, the obtained words are semantically-unrelated verb (CiCeC: [XiZeR] 'court.3sg.m.pst') and noun (CaCiC: [XaZiR] 'pig.m'). Moreover, the root . . /T.G.N/ produces an existing verb in the CiCeC pattern ( [TiGeN] 'fry.3sg.m.pst'), but a non-existing word in the CaCiC pattern (* [TaGiN]).

Hebrew also has an additional morpho-orthographic property that might be confusing for learners and also difficult for automatic parsers. Seven Hebrew function words consist of a single letter that is cliticized in the script to the following word (Fabri et al., 2014). These include the definite article /ha-/ 'the', the coordinating conjunction /ve-/ 'and', the relativizer /še-/ 'that', and the prepositions /be-/ 'in', /ke-/ 'as', /le-/ 'to' and /me-/ 'from'. Other function words stand alone in the script. Since all the cliticized letters can also be a part of a lexical unit (consonantal roots and root-less words), there are many morphologically ambiguous words in Hebrew (e.g., Share & Bar-On, 2018). For example, in the word the first letter can be either a part of the root (**Table 1**a) or a preposition (**Table 1**b). The exact reading is context-dependent.

**Table 1** Readings of the morphologically ambiguous .

| | | | | | |
|---|---|---|---|---|---|
| Root : | [lavan] | 'white.sg.m' | / | [lében] | 'Leben (dairy)' |
| Preposition : | [la-ben] | 'to the boy' | / | [le-bén] | 'to a boy'/'to Ben' |

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).[5] In particular, letter-transposition neighbors are relatively common in Hebrew (e.g., [to'elet] 'benefit' – [tola'at] 'worm' ; [xalav] 'milk' – [xevel] 'rope' / [xaval] 'a shame' ; [xazar] 'return.3sg.m.pst' – [xaraz] '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.

# 2.3 Morpho-syntax

The syntactic and morpho-syntacic 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, and adjectives. Since grammatical gender of inanimate objects is arbitrary,[6] 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 plural affix or correct verbal and adjectival forms to maintain agreement with the modified noun. This is especially challenging with nouns that have an irregular plural form. For example, masculine nouns are typically pluralized via the suffix /-im/ (e.g., [mispar] 'number' → [mispar-im] 'numbers'). When modified by an adjective, the adjective takes the plural form that matches the gender of the noun (e.g., [mispar-im gdol-im] 'large numbers'). However, certain masculine nouns are pluralized via the suffix /-ot/, which is typically used with feminine nouns. An adjective modifying such nouns takes the regular masculine suffix rather than the irregular suffix of the noun (e.g., [šulxan-ot gdol-im] 'big tables', not: * [šulxan-ot gdol-ot]).

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 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, /be/, and there is no simple mapping between uses of English and Hebrew prepositions (see also Hermet & Désilets, 2009).

[1] Uppercase letters represent root consonants, lowercase letters represent templatic vowels and consonants.
[2] Uppercase C represents any root consonant.
[3] 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*).
[4] For example, in French *table* 'table' is feminine, while the Hebrew equivalent [šulxan] is masculine.

# 3. The Corpus7

# 3.1 The essays

The corpus 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, a general test required for admission by most higher education institutions in Israel. The authors were either native speakers of Hebrew or candidates who decided to take the psychometric test in Hebrew even though they were not native speakers of Hebrew (the test is also administered in several other languages).[8] The essays by non-native speakers were collected as part of the *YAEL* test: 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. 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 the tests also differed between the two groups: the allotted time for essay writing was 15 minutes in the YAEL test and 30 minutes 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.

# 3.2 Metadata

The only available metadata for the native speaker essays is the essay score, in the range 1-6 (average: 3.67). 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.
- **Sex**: Male, Female, Unspecified.
- **Age**: 13-50 (average: 21).
- **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 (average: 20.7).[9] 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 (average: 540).
- **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 2** summarizes the average number of sentences and tokens per essay in each of the three L1s. The average number of sentences per essay in the native sub-corpus was 15.2 (SD: 5.3), and the average 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 Section 3.1).

**Table 2** Average numbers of sentences and tokens per essay across L1s in the non-native corpus. Numbers in parentheses denote standard deviation.[10]

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

**Fig. 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.

**Fig. 2** depicts the average 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 significantly 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).

## 3.3 Processing

The essays, originally hand-written, were transcribed by NITE and stored in text files. 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 3** illustrates the processed representation of sentence (1) and its revision (1').

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

| ve-kaxa | **ha** | **ce'irim** | **yaxlu** | lilmod | bidiyuk | axrey | **le-siyum** | beit | sefer |
|---|---|---|---|---|---|---|---|---|---|
| and-this way | **the** | **young.m.pl** | **could.pl** | study.inf | exactly | after | **to-end** | house.constr | book |

(1') □□□□ □□□□□□□□□□□□

| ve-kaxa | **ha-ce'irim** | **yuxlu** | lilmod | bidiyuk | axrey | **siyum** | beit | sefer |
|---|---|---|---|---|---|---|---|---|
| and-this way | **the-young.m.pl** | **can.pl** | study.inf | exactly | after | **end** | house.constr | book |

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

**Table 3** A processed text

| | Token | TH1 |
|---|---|---|
| 1 | ve-kaxa | ve-kaxa |
| 2 | **ha** | **&&** |
| 3 | **ce'irim** | **ha-ce'irim** |
| 4 | **yaxlu** | **yuxlu** |
| 5 | lilmod | lilmod |
| 6 | bidiyuk | bidiyuk |
| 7 | axrey | axrey |
| 8 | **le-siyum** | **siyum** |
| 9 | beit | beit |
| 10 | sefer | sefer |

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 3** to maintain alignment between the revised token [ha-ce'irim] 'the-young.m.pl' and the corresponding split token in the original text [ha ce'irim] 'the young.M. PL').

[7] 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.

[8] We did not have access to information on the native language of these authors, and we therefore consider them all "natives".

[9] The full range of scores in the YAEL test is 4-28.
[10] 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 Section 5.

# 4. 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 Section 4.3). 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 4** 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).

**Table 4** Statistics of annotated non-native essays per L1

| L1 | Essays | Sentences | Tokens |
|---|---|---|---|
| Arabic | 342 | 2023 | 50304 |
| French | 338 | 2989 | 48893 |
| Russian | 333 | 2993 | 47213 |
| **Total** | 1013 | 8005 | 146410 |

Our annotations consists 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 (Section 4.1). Finally, we also offer interpretations pertaining to the presumed cause of these errors (formulated as a basic classification of the errors by type/cause) in Section 4.2.

# 4.1 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 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 Section 2).

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.

### 4.1.1 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.[11] For example, consider the following sentence:

(2) □□□□ *

| hu | carix | lehasig | **mašehu** | roce |
|---|---|---|---|---|
| he | need.sg.m.prs | achieve.inf | **something** | want.sg.m.prs |

*'He needs to achieve something wants'

(2) is ungrammatical. The most conservative interpretation would be to treat [mašehu] 'something' as a morpho-orthographic error, an incorrect merging of the words [ma še-hu] 'what that-he'. The hypothesized target sentence is then:

(2') □□□□ □□

| hu | carix | lehasig | **ma** | **še-hu** | roce |
|---|---|---|---|---|---|
| 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 Section 4.2.

### 4.1.2 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).[12] 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 Section 2.2), 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 examples illustrate the application of the cooperative principle:

(3) □□□□□□

| ze | yavo | rak | **le-tola'at** | ha-mišpaxa | acma |
|---|---|---|---|---|---|
| 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: __ 'to the worm of' should probably have been __ 'to the benefit of'. We annotate this as a spelling error and introduce a correction. The hypothesized target sentence is:

(3') □□□□□□

| ze | yavo | rak | **le-to'elet** | ha-mišpaxa | acma |
|---|---|---|---|---|---|
| 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'

### 4.1.3 The faithfulness principle: Minimal editing and information maximization

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 /maxšev/ 'computer', but there are several suitable alternatives, including /be-/ 'in', /mi-/ 'from', and /be-emca'ut/ 'using'. In this case, the first two alternatives are preferable, since the prepositions and are used as clitics and, therefore, do not affect tokenization. This is demonstrated in (4'). By contrast, adding the stand-alone preposition is dispreferred, since it increases the number of words in the text (see 4").

(4) □□□□ *

| hayom | efšar | likro | ma | kore | be-sin | **maxšev** | še-nimca | be-pariz |
|---|---|---|---|---|---|---|---|---|
| today | possible | read.inf | what | happen.sg.m.prs | in-China | **computer** | that-situated.sg.m.prs | in-Paris |

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

(4’) □□□□□

| hayom | efšar | likro | ma | kore | be-sin | **mi-maxšev** | še-nimca | be-pariz |
|---|---|---|---|---|---|---|---|---|
| today | possible | read.inf | what | happen.sg.m.prs | in-China | **from-computer** | that-situated.sg.m.prs | in-Paris |

‘Today it is possible to read what happens in China from a computer located in Paris’

(4”) □□□□ □□□□□□□

| hayom | efšar | likro | ma | kore | be-sin | **be-emca’ut** | **maxšev** | še-nimca |
|---|---|---|---|---|---|---|---|---|
| today | possible | read.inf | what | happen.sg.m.prs | in-China | **using** | **computer** | that-situated.sg.m.prs |

be-pariz

in-Paris

‘Today it is possible to read what happens in China using a computer located in Paris’

According to the “information maximization” 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 non-linear 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) □□□□□□ *

| **lehaspik** | lahem | et | kol | craxeyhem |
|---|---|---|---|---|
| suffice.inf | to.them | acc | all | needs.poss.3pl.m |

*‘To suffice them all their needs’

(5) is ungrammatical due to a mismatch between the verb and its arguments. The verb [lehaspik] ‘suffice’ is assigned two internal arguments here: [lahem] ‘to them’ and [et kol craxeyhem] ‘all their needs’. Of the two arguments, only the first fits into the argument structure of the verb.[13] However, omitting the second argument will lead to a loss

of information. Furthermore, the resulting phrase will still be ungrammatical (or at least odd) in the original wider context:

(5’) □□□□□□ ???

| yeš | horim | še-lahem | eyn | maspik | kesef | kedey | **lehaspik** | lahem |
|---|---|---|---|---|---|---|---|---|
| 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 /lehaSPiK/ to a verb of the same root in a different Binyan (verb pattern): /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”) □□□□

| **lesapek** | lahem | et | kol | craxeyhem |
|---|---|---|---|---|
| provide.inf | to.them | acc | all | need.pl.poss.3pl.m |

‘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 [latet] ‘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’”) □□□

| **latet** | lahem | et | kol | craxeyhem |
|---|---|---|---|---|
| give.inf | to.them | acc | all | need.pl.poss.3pl.m |

‘To give them all their needs’

## 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”).

???

| af | exad | lo | mistakel | al | ha-axer | o | noten | lo | et | ha-inyan |
|---|---|---|---|---|---|---|---|---|---|---|
| no | one | neg | look.sg.m.prs | on | the-other | or | give.sg.m.prs | 3sg.m.dat | acc | the-interest |

??? ‘No one looks at the other or gives him the interest’

(6’)

| af | exad | lo | mistakel | al | ha-axer | o | noten | lo | et |
|---|---|---|---|---|---|---|---|---|---|
| no | one | neg | look.sg.m.prs | on | the-other | or | give.sg.m.prs | 3sg.m.dat | acc |

| tsumet | ha-lev |
|---|---|
| input.constr | the-heart |

‘No one looks at the other or gives attention to them’

(6”)

| af | exad | lo | mistakel | al | ha-axer | o | mitanyen | bo |
|---|---|---|---|---|---|---|---|---|
| no | one | neg | look.sg.m.prs | on | the-other | or | take-interest.sg.m.prs | 3sg.m.loc |

‘No one looks at the other or takes interest in them’

In the Hebrew Essay Corpus, multiple target hypotheses are indicated in separate columns, as shown in **Table 5**. 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.

**Table 5** Multiple target hypotheses

| **Token** | **TH1** | **TH2** |
|---|---|---|
| af | af | |
| exad | exad | |
| lo | lo | |
| mistakel | mistakel | |
| al | al | |
| ha-axer | ha-axer | |
| o | o | |
| noten | noten | mitanyen |
| lo | lo | bo |
| et | et | && |
| ha-inyan | **tsumet** | && |
| && | **ha-lev** | && |

Another kind of uncertainty occurs when the intended meaning is unclear. In such cases, annotators were advised to leave the text uncorrected and, instead, make free-form comments, or assign special error tags to parts of the text during the error annotation process (see Section 4.2.4). For example, consider the following sentence:

(7) □□□□ *

| nir'e | li | še-yeš | mašehu | še-dome | kmo | **haxala** | texnologya |
|---|---|---|---|---|---|---|---|
| seem.sg.m.prs | to.me | that-exist | something | that-similar.sg.m | like | **application** | technology |

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

The phrase " " [haxala texnologya] 'technology application' is ungrammatical, but it is not clear what the intended meaning was (if the author meant 'application of technology' it seems that some information is missing, e.g., 'application to what'?). In fact, it is not clear at all that the author meant to use the word " " 'application', 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 /dome/ '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 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.

# 4.2 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 6** 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 misspells (e.g., *teknology*) that mirror orthographic errors in the Hebrew text.

(8) *

| yoter | ha-texnologya | mitpateax | yoter | kaše | ze | limco | mašehu |
|---|---|---|---|---|---|---|---|
| more | ~the-teknology | evolve.sg.m.prs | more | hard.sg.m | this | find.inf | something |

| še-lo | mistakel | et | ha-telefon | kol | daka |
|---|---|---|---|---|---|
| that-neg | look.m.sg.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’)

| kexol | še-ha-texnologya | mitpataxat | yoter | kaše | limco | mišehu |
|---|---|---|---|---|---|---|
| as much | the-technology | evolve.sg.f.prs | more | hard.sg.m | find.inf | someone |

| še-lo | mistakel | ba-telefon | kol | daka |
|---|---|---|---|---|
| that-neg | look.m.sg.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’

**Table 6** Tokenized, revised and annotated text

| | Token | TH1 | Error1_TH1 | Error2_TH1 |
|---|---|---|---|---|
| 1 | yoter | kexol | wrong(conj) | |
| 2 | ha-texnologya | še-ha-texnologya | shouldB( , ) | miss(conj,##) |
| 3 | mitpateax | mitpataxat | agree(subj,pred) | |
| 4 | yoter | yoter | | |
| 5 | kaše | kaše | | |
| 6 | ze | && | redun(dem) | |
| 7 | limco | limco | | |
| 8 | mašehu | mišehu | oMiss( ) | |
| 9 | še-lo | še-lo | | |
| 10 | mistakel | mistakel | | |
| 11 | et | **&&** | wrong(prep,&&) | |
| 12 | ha-telefon | ba-telefon | wrong(prep) | |
| 13 | kol | kol | | |
| 14 | daka | daka | | |

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

### 4.2.1 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 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 6** 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 categories of the two elements for which there is a lack of agreement in gender, number, or person. For instance, row 3 in **Table 6** contains the tag agree(subj,pred), indicating an agreement error between the feminine subject of the clause, [texnologya] 'technology' and the main predicate of the clause [mitpateax] 'evolve.sg.m.prs', which is masculine.

### 4.2.2 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 , which does not exist in Hebrew. In (9'), it was corrected to [yexolim] 'can.m.pl.prs', resulting in a grammatical sentence.

(9) □□□□□□ *

| ha-ce'irim | lo | **yoxlim** | la'avod |
|---|---|---|---|
| the-young.m.pl | neg | ~abel | work.inf |

(9') □□□□□□

| ha-ce'irim | lo | **yexolim** | la'avod |
|---|---|---|---|
| the-young.m.pl | neg | able.m.pl.prs | work.inf |

'The young are unable to work'

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., → ). 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 7** demonstrates alternative analyses of the same error in the Hebrew Essay Corpus. The TH1 column contains the full revised text, as explained earlier. The TH2 column contains a copy of the revised token [yexolim] (this is in contrast to situations described in Section 4.1.4, in which TH2 was different from TH1). Alternative analyses of the error are indicated in the Error1_TH1 and Error1_TH2 columns.

**Table 7** Alternative error analyses

| Token | TH1 | Error1_TH1 | TH2 | Error1_TH2 |
|---|---|---|---|---|
| ha-ce'irim | ha-ce'irim | | | |
| lo | lo | | | |
| **yoxlim** | **yexolim** | metathesis( ) | yexolim | wrong(pattern) |
| la'avod | la'avod | | | |

# 4.2.3 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 Section 3.3, 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 will 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 8** demonstrates the annotation of a dummy token inserted as part of a preposition correction in sentence (8). The correction replaced the stand-alone preposition [et] (an accusative marker) by the cliticized preposition 'in'. Overall, the single preposition correction resulted in the change of two tokens. The change in row 9 was tagged wrong (prep), while 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.

**Table 8** Error tags and dummy tokens

| | Token | TH1 | Error1_TH1 |
|---|---|---|---|
| 1 | yoter | yoter | |
| 2 | kaše | kaše | |
| 3 | ze | && | redun(dem) |
| 4 | limco | limco | |
| 5 | mašehu | mišehu | oMiss( ) |
| 6 | še-lo | še-lo | |
| 7 | mistakel | mistakel | |
| 8 | et | **&&** | wrong(prep,&&) |
| 9 | ha-telefon | ba-telefon | wrong(prep) |
| 10 | kol | kol | |
| 11 | daka | daka | |

Note that not all dummy tokens were tagged with &&. Row 3 in **Table 8** demonstrates deletion of the demonstrative [ze] '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.

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 the Hebrew Essay Corpus errors by ## as an argument to the 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 /be/ ‘in’ is repeated instead of the preposition /le/ ‘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 9**).

(10) □ □ *

| anašim | rocim | lehacliax | ba-xaim |
|---|---|---|---|
| people | want.pl.m.prs | succeed.inf | in.the-life |

| ve-samim | lev | yoter | **ba**-avoda | ve-lo | **ba**-mišpaxa |
|---|---|---|---|---|---|
| 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’) □ □

| anašim | rocim | lehacliax | ba-xaim |
|---|---|---|---|
| people | want.pl.m.prs | succeed.inf | in.the-life |

| ve-samim | lev | yoter | **la**-avoda | ve-lo | **la**-mišpaxa |
|---|---|---|---|---|---|
| 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’

**Table 9** Annotation of a “chain correction” in a coordination construction

| Token | TH1 | Error1_ TH1 |
|---|---|---|
| ve-samim | ve-samim | |
| lev | lev | |
| yoter | yoter | |
| **ba**-avoda | **ba**-avoda | wrong(prep) |
| ve-lo | ve-lo | |
| **ba**-mišpaxa | **ba**-mišpaxa | wrong(prep,##) |

Another type of chain correction involves reattachment of clitics. Recall that Hebrew has several function words that are attached as clitics to the following word. 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 /ha-š'ar dvarim/ '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 10**, to avoid inflating the number of estimated errors in the corpus.

(1) □□□□□ □□□□ *

| letapel | bexol | **ha-š'ar** | **dvarim** | ba-bait |
|---|---|---|---|---|
| handle.inf | in-all | **the-rest** | **things** | at.the-house |

(11') □□□□□□ □□□

| letapel | bexol | **š'ar** | **ha-dvarim** | ba-bait |
|---|---|---|---|---|
| handle.inf | in-all | **rest** | **the-things** | at.the-house |

'To take care of the rest of the stuff at home'

**Table 10** Annotation of clitic reattachment

| Token | TH1 | Error1_ TH1 |
|---|---|---|
| ha-š'ar | š'ar | redun(det) |
| dvarim | ha-dvarim | miss(det,##) |

# 4.2.4 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 deontic element, such as [adif] '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 11**.

(12) *

| leda'ati | lehikanes | le-nose | še-yoter | kal | lexa/lax |
|---|---|---|---|---|---|
| in-my-opinion | enter.inf | to-subject | that-more | easy.sg.m | to-you |

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

(12') □□□□

| leda'ati | **adif** | lehikanes | le-nose | še-yoter | kal | lexa/lax |
|---|---|---|---|---|---|---|
| in-my-opinion | **preferable** | enter.inf | to-subject | that-more | easy.sg.m | to-you |

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

**Table 11** Error tagging of an unknown missing lexical item

| Token | TH1 | Error1_ TH1 |
|---|---|---|
| leda'ati | leda'ati | |
| && | && | miss(lex,$$) |
| lehikanes | lehikanes | |

## 4.2.5 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 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 12** 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).

The use of [kaxa] 'this way' instead of [ze] 'this' is analyzed as a wrong demonstrative (row 2). In addition, it can be viewed as a miscollocation – deformation of the collocation [biglal ze] 'because of that'. The use of [ya'ase] 'do.3sg.fut' instead of [e'ese] '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. Finally, [bsixometri] '~bsychomettric (test)' exhibits two spelling errors (row 5). The - 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 ) 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 - substitution is a homophonic letter substitution (both letters represent the consonant /t/).

(13) *

| biglal | kaxa | ani | ya'ase | bsixomeTri |
|---|---|---|---|---|
| because-of | this-way | I | do.3sg.m.fut | ~bsychomettric (test) |

*'Because of this way I will take (3sg.m) the bsychomettric (test)'

(13')

| biglal | ze | ani | e'ese | psixometri |
|---|---|---|---|---|
| because-of | this/that | I | do.1sg.fut | psychometric (test) |

'Because of that I will take the psychometric (test)'

**Table 12** Annotated text with explanatory error tags

| | Token | TH1 | Error1_TH1 | Interp1_TH1 | Error2_TH1 | Interp2_TH1 |
|---|---|---|---|---|---|---|
| 1 | biglal | biglal | | | | |
| 2 | kaxa | ze | wrong(dem) | colloc | | |
| 3 | ani | ani | | | | |
| 4 | ya’ase | e’ese | shouldB( , ) | pronuncReg /register | | |
| 5 | bsixomeTri | psixometri | shouldB( , ) | pronuncL2 | shouldB( , ) | 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), pronuncL2 = pronunciation of L2 speakers (with a specific L1), homophone = homophonic letter substitution

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.

## 4.3 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% the size of the annotated corpus. The number of words corrected by both annotators was 667, 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 13**.

**Table 13** Categories of disagreements on corrections

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

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 4.1.4 above). For example, the phrase in (14) is ungrammatical. Both annotators corrected the word [ha-rišon] 'the-first.sg.m', but each applied a different but equally acceptable correction (see 14' and 14").

(14) *

| Matrat | **ha-rišon** | šel | ha-aplikacyot |
|---|---|---|---|
| goal (f).constr | **the-first.sg.m** | of | the-applications |

*'The goal(f) of first(m) of the applications'

(14')

| ha-matara | **ha-rišona** | šel | ha-aplikacyot |
|---|---|---|---|
| the-goal (f) | **the-first.sg.F** | of | the-applications |

'The first goal of the applications'

(14")

| ha-matara | **ha-mekorit** | šel | ha-aplikacyot |
|---|---|---|---|
| the-goal (f) | **the-original.sg.m** | 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 4.2.3, 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 [me] ‘from’, resulting in a difference in two tokens between the two corrections, as demonstrated in **Table 14**.

(15) *

| anašim | še-hit’abdu | me-atarey | internet |
|---|---|---|---|
| people | that-commit suicide.pl.pst | from-sites.constr | internet |

*‘People who committed suicide from websites’

(15’)

| anašim | še-hit’abdu | **ke-toca’a** | **me-atarey** | internet |
|---|---|---|---|---|
| people | that-commit suicide.pl.pst | **as-result** | **from-sites.constr** | internet |

‘People who committed suicide as a result of websites’

(15”)

| anašim | še-hit’abdu | **biglal** | **atarey** | internet |
|---|---|---|---|---|
| people | that-commit suicide.pl.pst | **because-of** | **sites.constr** | internet |

‘People who committed suicide because of websites’

**Table 14** Disagreement in chain corrections

| **Token** | **Annotator1** | **Annotator2** | |
|---|---|---|---|
| anašim | anašim | anašim | |
| še-hit’abdu | še-hit’abdu | še-hit’abdu | |
| **&&** | **ke-toca’a** | **biglal** | |
| **me-atarey** | **me-atarey** | **atarey** | ← chain correction |
| internet | internet | internet | |

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 [le] ‘to’ in (16) to the bound preposition [be] ‘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 ([be-davar] ‘in-thing’ vs. [be-mašehu] ‘in-something’). However, the fact that they did agree on the correction of the preposition should not be overlooked.

(16) *

| | | | |
|---|---|---|---|
| xafec | be-xol | libo | **le-davar** |
| wish.sg.m.prs | in-all | heart.poss.3sg.m | **to-thing** |

*‘Wishes with all his heart to a thing’

(16’)

| | | | |
|---|---|---|---|
| xafec | be-xol | libo | **be-davar** |
| wish.sg.m.prs | in-all | heart.poss.3sg.m | **in-thing** |

‘Wishes with all his heart for a thing’

(16’)

| | | | |
|---|---|---|---|
| xafec | be-xol | libo | **be-mašehu** |
| wish.sg.m.prs | in-all | heart.poss.3sg.m | **in-something** |

‘Wishes with all his heart for something’

To summarize, when analyzing learner texts that have been corrected, one should keep in mind that the corrections do not represent 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.

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 15**.

**Table 15** Categories of disagreements on annotations

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

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 bound functional morphemes, such as prepositions (see 2.2). Such errors could, in principle, be analyzed as orthographic errors or as errors in a function word. For example, in one case, both annotators corrected the word [karov] ‘close’ to

[bekarov] ‘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 disagreements includes cases in which none 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 4.2.4). 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 [ š] to [ iš] 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, and our annotation scheme and guidelines were designed with this fact 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. 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.

[11] 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).
[12] In the Hebrew essay corpus, 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”).
[13] The verb is ambiguous. One of its uses does take a direct object, but its meaning (‘to succeed doing something on time’) is incompatible with the given context.

# 5. Analysis

To demonstrate the utility of the corpus, we compared the number of errors in various linguistic categories across the three L1s (see **Table 16**). Each error tag was classified as belonging to one of the linguistic categories and we calculated 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).

**Table 16** 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., af = a significant difference between the distributions of a certain error type in essays authored by L1 Arabic and L1 French). Finally, the table specifies the total number of errors tagged for each L1.

**Table 16** Error tags across the three L1s

| Category | Arabic | French | Russian | H(2) | Sig |
|---|---|---|---|---|---|
| *Orthography* | | | | | |
| General | 2241 | 2944 | 3135 | 18.90 | *** ar, af |
| | | | | | |
| *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 | *** ar, af, fr |
| | | | | | |
| *Syntax* | | | | | |
| Agreement | 760 | 731 | 705 | 1.39 | |
| Argument structure | 182 | 129 | 118 | 14.76 | ** ar, af |
| Conjunctions | 753 | 485 | 372 | 99.24 | *** ar, af, fr |
| Construct state | 43 | 49 | 26 | 5.39 | |
| Copulas | 147 | 70 | 91 | 27.68 | *** ar, af |
| Pronouns and demonstratives | 159 | 133 | 96 | 11.50 | ** ar |
| Determiners | 322 | 389 | 925 | 158.11 | *** ar, fr |
| Existentials | 52 | 33 | 25 | 7.50 | * ar |
| Negation | 22 | 21 | 20 | 0.19 | |
| Order | 182 | 124 | 133 | 4.24 | |
| Prepositions | 943 | 1066 | 667 | 60.39 | *** ar, fr |
| 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 | *** ar, af |
| | | | | | |
| **Tokens** | 50304 | 48893 | 47213 | | |
| **Errors** | 8232 | 7907 | 7495 | | |

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

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

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.

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 2**).

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 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 the Hebrew Essay Corpus will be invaluable both for research in learner language, including for example transfer effects from L1, and for practical educational applications.

# 6. Conclusions

We presented the Hebrew Essay Corpus, 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 both the native and non-native sub-corpora. 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.

## Declarations

The authors have no relevant financial or non-financial interests to disclose.

# Data availability statement

All the data used in this research is available upon request. See footnote 7 and the accompanying datasheet.

# 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. 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.

We would also like to thank Gur Meir for his help with the error tagging process.

## References

1. 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*, 63–69.
2. 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
3. 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
4. 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
5. 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*, 327–337.

6. 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*, 94–102. https://doi.org/10.18653/v1/k15-1010
7. Bykh, S., & Meurers, D. (2012). Native Language Identification Using Recurring N-grams – Investigating Abstraction and Domain Dependence. *Proceedings of COLING 2012*, 425–440.
8. 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
9. 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
10. 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
11. 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*, 263–272.
12. 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
13. Frost, R. (2012). Towards a universal model of reading. *Behavioral and Brain Sciences*, *35*(05), 263–279. https://doi.org/10.1017/S0140525X11001841
14. 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). https://doi.org/ISBN 978-0-85261-941-4
15. Gafni, C., Prior, A., & Wintner, S. (2022). The Hebrew Essay Corpus. *Proceedings of the 13th Conference on Language Resources and Evaluation*, 5580–5586.
16. 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
17. 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/10.1163/9789401206204_002
18. Gilquin, G., & Paquot, M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction*, *1*(1), 41–61.
19. 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*, 3591–3601. https://doi.org/10.18653/v1/d18-1395
20. 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.
21. 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
22. 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.

23. 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
24. Grice, P. (1989). *Studies in the Way of Words*. Harvard University Press.
25. 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.
26. Gries, S. T. (2015). Statistics for learner corpus research. In *The Cambridge Handbook of Learner Corpus Research* (pp. 159–181). Cambridge University Press.
27. 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
28. 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*, 64–72. https://doi.org/10.3115/1609843.1609853
29. 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.
30. 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*, 624–628. https://doi.org/10.1145/1081870.1081947
31. Nguyen, I., & Wintner, S. (2022). Predicting the Proficiency Level of Nonnative Hebrew Authors. *Proceedings of the Language Resources and Evaluation Conference*, 5356–5365. https://aclanthology.org/2022.lrec-1.573
32. 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
33. 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).
34. 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
35. Schütze, C. T. (2016). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Language Science Press. https://doi.org/10.26530/oapen_603356
36. 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
37. 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
38. Tetreault, J., Blanchard, D., & Cahill, A. (2013). A Report on the First Native Language Identification Shared Task. *Aclweb.Org*, 48–57. http://www.aclweb.org/anthology/W13-1706
39. 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*, 1–8.
40. 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.

41. 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*, 279–287. https://doi.org/10.1001/archophthalmol.2010.205
42. 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

# Figures

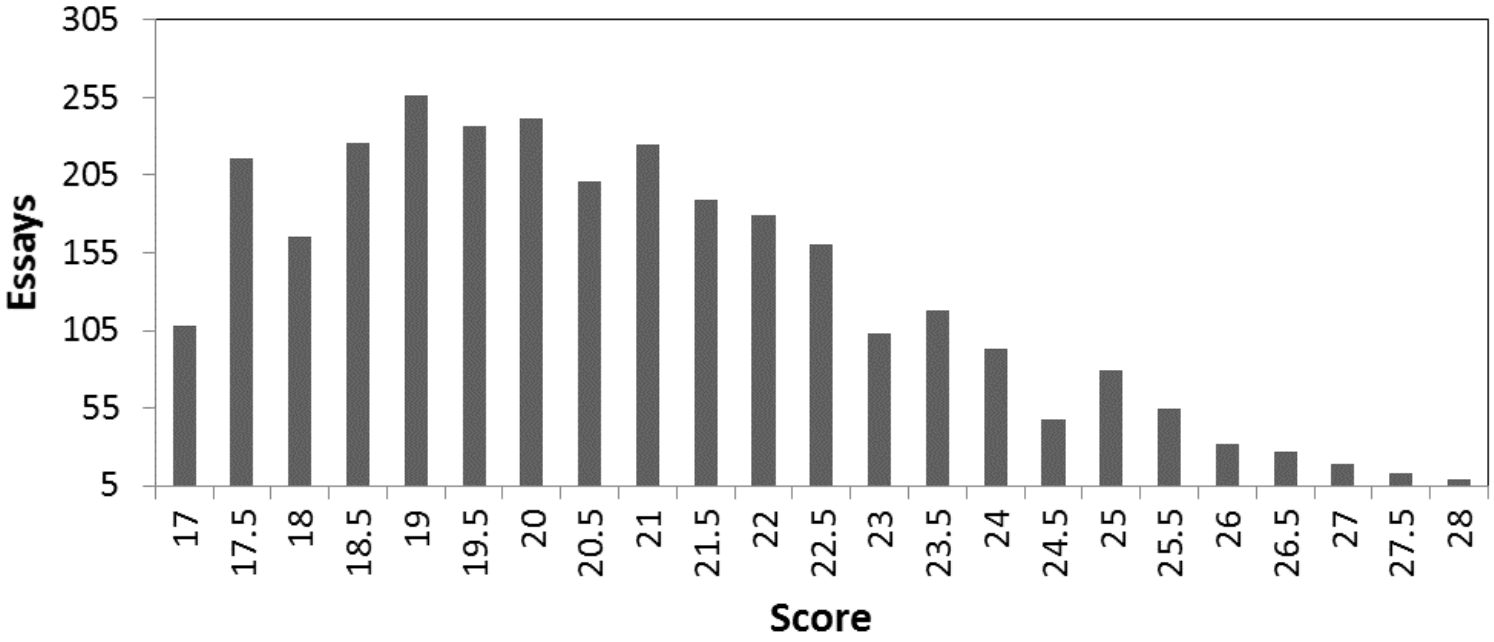


**Figure 1**

Distribution of essays by score

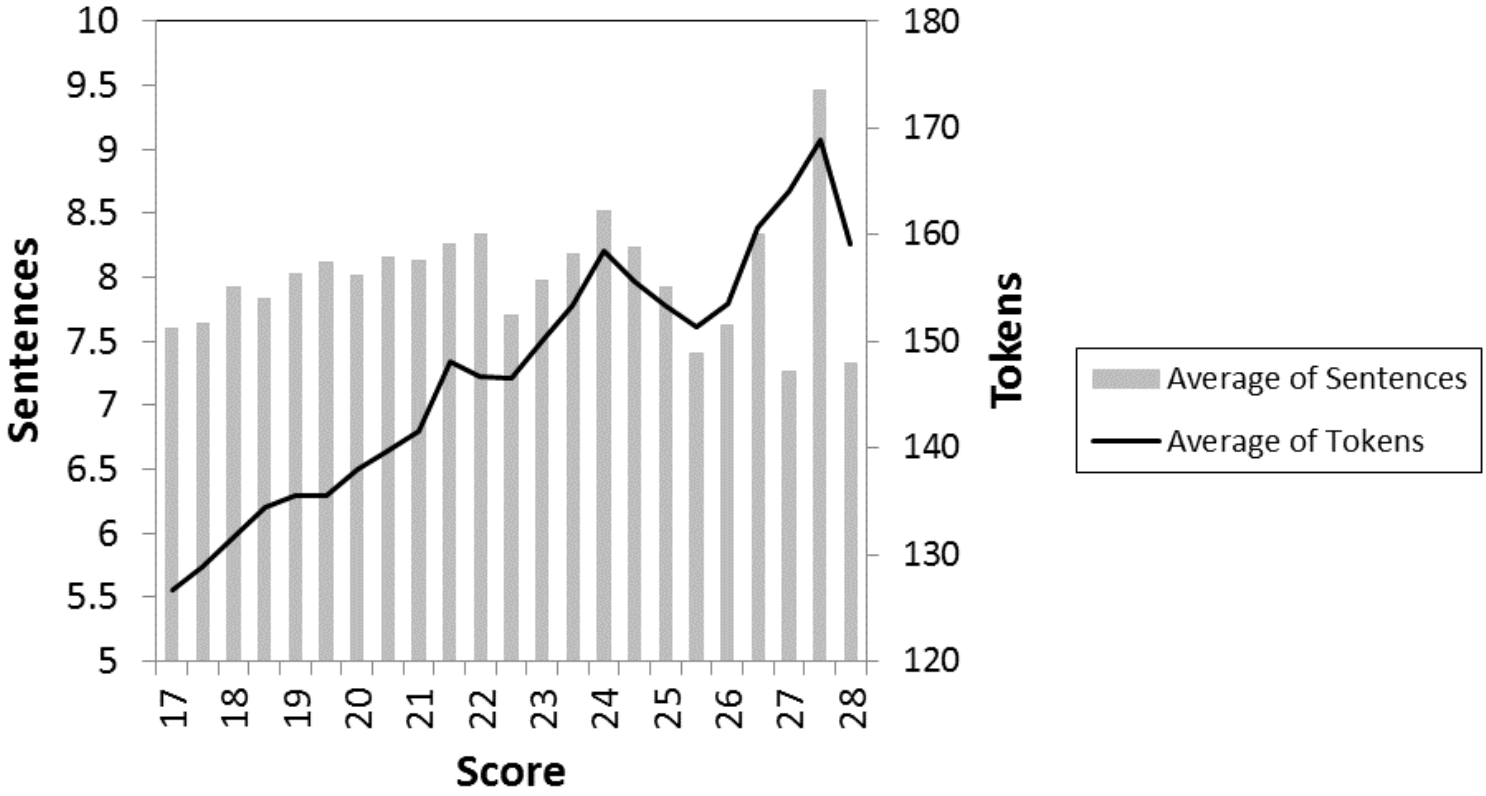


**Figure 2**

Average numbers of tokens and sentences per essay across YAEL test scores

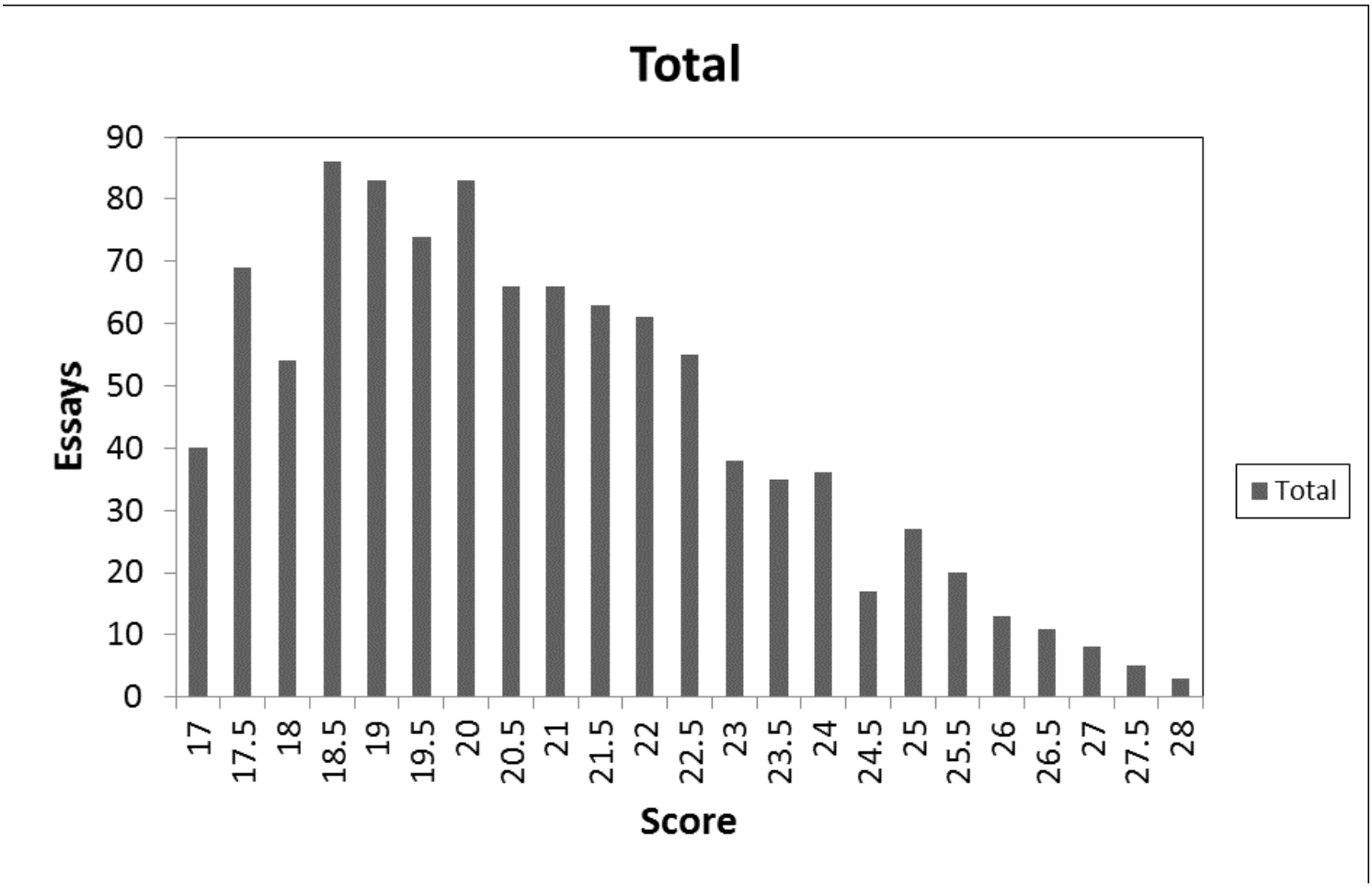


**Figure 3**

Number of annotated non-native essays per test score

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendixscheme.pdf
- Datasheet.pdf