# HelexKids 2.0: A linguistically annotated lexical database building on HelexKids

**Authors:**

Anthi Revithiadou, Department of Linguistics, School of Philology, Aristotle University of Thessaloniki, ORCID ID: 0000-0003-1432-1414

Aris Terzopoulos, Department of Psychology, Birmingham City University, ORCID ID: 0000-0003-3151-3805

Georgia Niolaki, School of Education, University of Birmingham, ORCID ID: 0000-0002-1825-141X

Giorgos Markopoulos, Department of Mediterranean Studies, University of the Aegean and School of Philology, University of Ioannina, ORCID ID: 0009-0000-3037-1425

Konstantinos Avdelidis, Department of Linguistics, School of Philology, Aristotle University of Thessaloniki, ORCID ID: 0009-0002-4940-8122

Ilias Mittas, Department of Linguistics, School of Philology, Aristotle University of Thessaloniki, ORCID ID: 0009-0000-1378-8560

Kosmas Kosmidis, Department of Physics, Aristotle University of Thessaloniki, ORCID ID: 0000-0002-9150-5746

**Corresponding Author:** Anthi Revithiadou, Department of Linguistics, School of Philology, Aristotle University of Thessaloniki, 54124 University Campus, Thessaloniki, Email: revith@lit.auth.gr

**Abstract**

HelexKids 2.0 represents an annotated expansion of the original HelexKids, a word frequency database compiled from Greek and Cypriot elementary school textbooks. This article presents the extended annotation methodology employed to enrich the database with part-of-speech tags, phonological properties (syllable count, phonetic transcription, orthographic and phonetic syllabification, and stress), and morphological features (case, number, gender, and inflection class for nouns). All original frequency and lexical measures are retained, and new ones (PLD20, bigram and biphone frequencies) are added. The annotation process combined automated tools with extensive manual annotation and expert verification. The database was redesigned to be dynamically assembled from the full set of textbooks, with results delivered on the fly. The annotated database comprises 67,802 types and over 1.3 million tokens. We report on stress pattern distributions across major parts of speech and grade levels, show how stress is distributed by word class, and provide hierarchies of word-level syllabic templates. HelexKids 2.0 is freely available through the GRADIENCE project webpage and serves as a valuable resource for psycholinguistic and morphophonological research, as well as for educational and computational applications.

**Keywords:** lexical database, Greek, elementary school vocabulary, morphophonological annotation, educational resources

# HelexKids 2.0: A linguistically annotated lexical database building on HelexKids

## 1 Introduction

Word frequency databases have become valuable resources for psycholinguistic research, educational assessment, and computational linguistics. They provide researchers with reliable data on word distributions and frequencies, assist educators in selecting vocabulary appropriate for different age groups, and offer linguists insights into both systematic and exceptional patterns in language. While numerous frequency databases exist for adults across several languages (Brysbaert & New, 2009; Keuleers et al., 2010), child-specific databases are relatively rare. This gap is significant because lexical exposure differs substantially between adults and children, a difference that proves important when selecting appropriate vocabulary items for teaching or when designing research materials (Beckman & Edwards, 2000; Massaro, 2015; Montag & MacDonald, 2015; Storkel, 2004).

For Greek, *HelexKids* (Terzopoulos et al., 2017; https://www.helexkids.org) was developed as the first psycholinguistic database specifically for primary school children. Based on official textbooks used across six grade levels in Greece and Cyprus primary schools, HelexKids offers frequency information that reflects the vocabulary children aged 6–12 are exposed to in the classroom. The database reveals several important characteristics of children's reading materials (Terzopoulos et al., 2017). First, approximately 50% of words at each grade level were hapax legomena, that is, words appearing only once in the total textbooks. Second, words were shorter and more frequent compared to adult Greek databases (e.g., *SUBTLEX-GR*, Dimitropoulou et al., 2010; *GreekLex 2.0*, Kyparissiadis et al., 2017). Finally, the top 100 words accounted for 44.9% of all tokens.

However, although the original *HelexKids* provided frequency and basic distributional metrics, it did not include grammatical information. For many research applications,

particularly in phonological development, morphophonological analysis, literacy acquisition, and the design of intervention materials for speech and writing difficulties, more detailed linguistic annotation is required (Terzopoulos et al., 2020). Greek, in particular, presents a range of challenges that underscore the need for fully annotated lexical databases, especially those targeting child learners. The language has morphology-dependent stress (see Alderete, 1999, 2001; Idsardi, 1992; Kiparsky, 1973, 1984; Kiparsky & Halle, 1977; Melvold, 1990; Revithiadou, 1999, among others), which entails that stress is not predictable; it can surface on any of the final three syllables (see Apoussidou, 2007; Burzio & Tantalou, 2007; Drachman & Malikouti-Drachman, 1999; Malikouti-Drachman & Drachman, 1989; Ralli & Touratzidis, 1991; Revithiadou, 1999, among others). Furthermore, the nominal system exhibits multiple inflectional classes (Anastassiadis-Symeonidis, 2012; Ralli, 2000, 2005) with distinct phonological and morphological properties that interact with stress placement (Apostolouda, 2018; Revithiadou & Lengeris, 2016; Revithiadou et al., 2026a, b; Sims, 2019).

Morphophonological properties also differ systematically between verbs and nominals. For instance, verb stress is generally more predictable and is determined by morphological features such as tense (with past tense forms typically being proparoxytone) or by conjugation class properties (non-past 1st person singular forms with a theme element display final stress, whereas those without a theme element are paroxytone). Verbs and nouns also differ in their phonotactic patterns: for example, the cluster /vɣ/, consisting of a voiced labio-dental and a dorsal fricative occurs word-initially only in verbs (e.g., [ˈvɣazo] ‘remove’), whereas the sequence /vð/ (e.g., [vðoˈmaða] ‘week’), consisting of a voiced labio-dental and a coronal fricative, occurs only in nouns.

Quantifying the presence of such systematic patterns, as well as the extent of idiosyncratic properties, is crucial, especially for the design of experimental materials and for the development of instructional and reading resources for young speakers (Korochkina et al.,

2024). The annotated version of HelexKids, named *HelexKids 2.0*, addresses exactly this gap for Greek. More specifically, it adopts a richer linguistic annotation scheme incorporating part-of-speech (PoS) tags, morphological features for nominal inflection, and phonological properties including detailed stress pattern information, phonetic (IPA) transcription, orthographic and phonemic syllabification, and word-level syllable templates. In this respect, the annotated HelexKids can also be used as an example for other languages since it offers a broad range of annotation parameters.

This article has three main objectives. First, we present the annotation methodology employed in HelexKids 2.0, including a detailed description of the semi-automated pipeline that combined existing resources, machine learning approaches, and principled manual annotation and verification. Second, we describe the database content including information on the distribution of major word classes as well as the distribution of phonological characteristics (stress, length, word-level syllable structure), per grade level or grade range. Third, we demonstrate the contribution of HelexKids 2.0 to theoretical research on the morphophonological development of Greek, and to literacy instruction and pedagogical practice.

## 2 Database annotation

In this section, we present the steps undertaken for the morphophonological annotation of HelexKids. The original lexical database comprises 76 textbooks and workbooks (some spanning multiple volumes) covering Greek language, arts, mathematics, science, history, and other subjects taught in elementary education (Grades 1–6) in Greece and Cyprus. The majority of these books are used exclusively at a single grade, with 16 shared across two or more grades.[1]

[1] Seven books are used exclusively in Grade 1, three are shared between Grades 1 and 2, and seven are exclusive to Grade 2. Nine books appear only in Grade 3, six are shared between Grades 3 and 4, and one spans Grades 4,

The textbooks were developed starting in 2007, officially approved in 2013, and are still in use. These books provide the 67,802 annotated word entries (types) of HelexKids 2.0 across six grades, and over 1.3 million tokens.

The database preserves the grade-level organization of the original HelexKids, with separate lexicons for each grade level (1–6) and cumulative grade ranges (Grades 1–2, 1–3, 1–4, 1–5, and 1–6). In addition, HelexKids 2.0 can generate lexicons for other grade combinations (e.g., Grades 2–3, 2–4, 3–5). It should be noted that, to reflect the actual material a student encounters, for books shared across grades, type and token counts in cumulative grades are assigned to the grade where each book is first introduced, and are not recounted in later grades in which the same book appears.

Each entry includes all annotations described in Sections 2.2–2.5: PoS category, syllable count, orthographic syllabification, IPA transcription and syllabification, word-level syllable templates, stress pattern, and morphological features (nouns only).

In addition to the morphophonological annotations, HelexKids 2.0 retains all variables from the original database, including raw frequency counts, estimated frequencies per million words (u), Zipf scale values (van Heuven et al., 2014), Orthographic Levenshtein Distance 20 (OLD20; Yarkoni et al., 2008),[2] Standard Frequency Index (SFI) scores, dispersion measures

---

5, and 6. Nine books are exclusive to Grade 4, fourteen to Grade 5, six are shared between Grades 5 and 6, and fourteen are exclusive to Grade 6.

[2] In HelexKids 2.0, the Zipf and OLD20 values from the original database have been recalculated due to minor differences in the total number of lexical items between the two databases. OLD20 was computed for each word type as the mean Levenshtein distance to its 20 closest orthographic neighbors across the full lexicon, using accent-sensitive NFC-normalized lowercase strings with punctuation removed and final sigma (ς) mapped to sigma (σ). The lowercase normalization follows standard practice in OLD20 computation (Terzopoulos et al., 2017). The final sigma mapping was not morphologically motivated. In Greek orthography, final sigma is simply the word-final form of sigma rather than a distinct letter, and the distinction carries no morpholexical significance.

(D), Contextual Diversity (CD) scores (Adelman et al., 2006), and word length in letters. Phonological Levenshtein Distance 20 (PLD20; Suárez et al., 2011; Yap & Balota, 2009) and bigram and biphone frequency (type and token; Vitevitch & Luce, 1999) were also added in the updated database.

HelexKids 2.0 differs from the original HelexKids in one crucial respect: it has been radically redesigned to be dynamically assembled on demand. Lexical items are extracted from the full set of textbooks, stored morphophonological information is retrieved, and additional metrics (e.g., SFI, D, CD), are computed on the fly (see Section 3 for details on the tool's architecture).

### 2.1 Pre-processing: Correction of character encoding

Before annotation could proceed, a systematic correction of the HelexKids Excel files was necessary. During the original compilation process, English letters had been accidentally substituted for visually similar Greek characters (e.g., Latin 'A' for Greek 'A', Latin 'o' for Greek 'o'). This problem was traced to the conversion of scanned PDF files to Excel format, which introduced character recognition errors. All instances were systematically identified and corrected to ensure proper recognition of Greek words by the annotation tools used in subsequent stages. A master Excel file containing all words in HelexKids was also prepared, which formed the basis for the annotation process.

### 2.2 PoS annotation

Collapsing the two therefore entails no loss of linguistically meaningful orthographic information. Without this mapping, the final sigma would incur an artificial substitution penalty in the word-final position. This would distort neighbourhood composition, making it reflect typographic convention rather than lexical similarity.

PoS annotation proceeded through multiple stages, combining existing resources with machine learning approaches, followed by extensive manual annotation and verification. Initial annotation used two existing annotated databases: GreekLex 2 (Kyparissiadis et al., 2017), which covers 35,282 types, and *A-Clean* (Apostolouda, 2018), a manually annotated database with PoS and morphosyntactic features that includes 72,103 types. It was extracted from the *Clean Corpus* (Protopapas et al., 2012), which in turn was derived from a previous version of the *Hellenic National Corpus* (ILSP, 2015; https://hnc.ilsp.gr). Automatic matching of these resources using exact word-form comparison reduced the number of untagged items. After this initial matching phase, approximately 45,000 words remained unannotated, representing about two-thirds of the total types in HelexKids 2.0.

For these words, we compiled a file of non-annotated lexical items and processed it using SpaCy's PoS tagging models for Greek. We tested both the large and the small models to optimize accuracy. However, SpaCy significantly underperformed on this task. For instance, certain words received incorrect tags: *τσάντα* [ˈʦanda] 'bag' was tagged as a proper noun rather than a common noun, and *απελπισία* [apelpiˈsia] 'despair' was incorrectly categorized as a verb.

Another challenge concerned forms that are morphosyntactically ambiguous between categories. For example, words ending in *-menos* (*-μενος*) are traditionally categorized as participles, yet many such forms function clearly as adjectives (e.g., *δυστυχισμένος* [ðistiçiˈzmenos] 'unhappy') or even nouns (e.g., *εργαζόμενη* [eryaˈzomeni] 'employed/employee') rather than as genuine participles (e.g., *κουρασμένος* [kuɾaˈzmenos] 'tired'). One criterion for identifying passive participles is their ability to form passive perfective constructions, a diagnostic that requires human judgment to apply.

Additionally, since the original HelexKids lacked context, words with syncretism created ambiguity. For instance, the homograph *διορθώσεις* [ðjor.ˈθo.sis] can mean either

‘correction’ (noun, plural) or ‘you correct’ (verb, 2nd person singular perfective). Moreover, several Greek words function across multiple grammatical categories depending on context: *ένα* [ˈena] can serve as a numeral (‘one’), an indefinite article (‘a/an’), or a pronoun (‘someone/something’). Finally, certain lexical items, such as *μπα* [ˈba] (‘nah, no way, I doubt it’) and *μακάρι* [maˈkaɾi] (‘if only, I wish’), cannot be easily assigned to a specific grammatical category and therefore require human judgment.

For all these reasons, we opted to resort to manual PoS annotation of the entire list of lexical items, based on specific, linguistically motivated criteria established by our research team. These criteria and the methodology applied are discussed in the following section.

### 2.3 Development of PoS tag list

Because the lexical items in HelexKids lack linguistic context, we developed a comprehensive PoS annotation guide to address those lexical items that were ambiguous with regard to their grammatical category. For this purpose, we relied on authoritative Greek grammatical references including Triantafyllidis (2018 [1941]), Holton et al. (2012), the *Dictionary of Standard Modern Greek* (1998), and the *Practical Dictionary of Modern Greek* (Charalambakis, 2023). The annotation scheme we developed uses compound tags (e.g., ADV/NOUN, NOUN/PNOUN, PCP/ADJ/NOUN) for the lexical items that belong to multiple categories. Table 1 presents representative examples of the simple and compound morphosyntactic categories employed in the annotation.

**Table 1**

*Representative categories of PoS annotation*

| Word | PoS | IPA transcription | Gloss |
|---|---|---|---|
| *ευθεία* | ADV/NOUN | [ef.ˈθi.a.] | ‘straight/line’ |

| | | | |
|---|---|---|---|
| *βιβλίο* | NOUN | [vi.ˈvli.o.] | ‘book’ |
| *παράγοντας* | NOUN/GER | [pa.ˈɾa.ɣo.ⁿdas.] | ‘factor/producing’ |
| *Ειρήνη* | NOUN/PNOUN | [i.ˈɾi.ni.] | ‘peace/Irene’ |
| *εργαζόμενη* | PCP/ADJ/NOUN | [er.ɣa.ˈzo.me.ni.] | ‘employed/employee’ |
| *εκφράσεις* | VERB/NOUN | [ek.ˈfra.sis.] | ‘(you-sg) express/ expressions’ |

Following the initial automated tagging attempts, manual annotation and correction were conducted by members of the research team. We decided that, when genuine ambiguity existed (i.e., when a word could legitimately function in multiple grammatical categories and context was insufficient to disambiguate), compound tags were required to capture this multifunctionality.

To be certain that the annotation scheme was applied consistently and reliably, we initially implemented a quality control process for PoS annotation consistency. A random sample of 1,000 words was independently annotated by three researchers. Inter-rater reliability was assessed using Krippendorff’s $\alpha$ (Krippendorff, 2004), which yielded $\alpha = 0.789$. This value falls below the threshold for reliable data ($\alpha \geq 0.800$) but exceeds the minimum for tentative conclusions ($\alpha \geq 0.667$). Disagreements were resolved through discussion and consultation with the team’s linguists and authoritative reference works. Most disagreements pertained to whether all applicable PoS tags had been included, rather than to misclassification of PoS category itself.[3]

[3] Words that were difficult to categorize (e.g., *Άη* [ˈaj.] ‘saint’), archaic (e.g., *Αδελφότης* [a.ðel.ˈfo.tis] ‘brotherhood’, *Επαναστάται* [e.pa.na.ˈsta.te] ‘revolutionaries’) or Ancient Greek words (e.g., *Αριστεύειν* [a.ri.ˈste.vin] ‘to excel’, *δικαίοις* [ði.ˈce.is] ‘to/for the righteous’), dialectal forms (e.g., *έφκιασε* [ˈef.ca.se] ‘s/he made’, *ήβρανε* [ˈi.vra.ne] ‘they found’), lexicographically unattested or non-existent items and invented words

As a result, annotation guidelines highlighting points of difficulty were developed and shared with the annotators. The annotation scheme contained 61 distinct PoS categories and category combinations. (See Appendix C, Table C1 for all categories in HelexKids 2.0.) The corrected annotations were then reinserted into the master Excel file, completing the PoS annotation of HelexKids 2.0.

### 2.4 Phonological annotation

Words in HelexKids 2.0 were annotated for five phonological properties: number of syllables, orthographic syllabification, IPA transcription and syllabification, stress pattern, and word-level syllable templates. All phonological annotations required manual checking to ensure accuracy.

#### 2.4.1 Orthographic and phonological syllabification

Orthographic syllabification was accomplished through a custom R script (Version 4.4.0; R Core Team, 2024) following Greek syllabification conventions (Petrounias, 2002).[4] The script implements standard rules for consonant cluster division, treating digraphs (e.g., *μπ, ντ, γκ*) as units and applying the *Maximum Onset Principle* (Kahn, 1976; Selkirk, 1982), whereby consonants are assigned to the onset of the following syllable whenever permitted by the phonotactic constraints of the language.

---

(e.g., *Αβγατηγανιστάν* [av.ɣa.ti.ɣa.ni.ˈstan], *έφκιακ* [ˈef.ci.ak]), and words containing orthographic or typographic errors (e.g., *εξαακολουθούν* [e.ksa.a.ko.lu.ˈθun] instead of *εξακολουθούν* [e.ksa.ko.lu.ˈθun] 'they continue', *τηλεγράφηματα* [ti.le.ˈɣra.fi.ma.ta] instead of *τηλεγραφήματα* [ti.le.ɣra.ˈfi.ma.ta] 'telegrams') were assigned to the category 'OTHER'.

[4] All scripts are openly accessible on the GRADIENCE webpage: https://gradience.lit.auth.gr/index.php/helexkids_2-0/

However, the orthographic syllabification applied to HelexKids 2.0 differs from traditional syllable division rulesin several crucial respects. First, it treats vowel sequences such as *ια, ιά, εια, ειά, ιο, ιό,* etc. as one syllable, depending on the lexical entry. Phonemically, these sequences are realized either as glides or palatal consonants within a single syllable or as independent syllables (see, e.g., Apostolopoulou, 2018; Baltazani et al., 2016; Baltazani & Topintzi, 2012; Markopoulos & Apostolopoulou, 2020; Rytting, 2005; Topintzi & Baltazani, 2016). This is subject to the particular lexical item and the learned/non-learned vocabulary distinction (Anastassiadis-Symeonidis & Fliatouras, 2004, 2018; Fliatouras & Anastassiadis-Symeonidis, 2019; Rytting, 2005). In examples like *μά.τια* [ˈma.tça] 'eyes'and *α.λή.θεια* [a.ˈli.θça] 'truth', the orthographic *ι* and *ει* correspond to a palatal consonant. However, in *διαφωνείτε* [ðja.fo.ˈni.te] 'you (pl) disagree' *ι* corresponds phonetically to a glide, whereas in *διαγράφω* [ði.a.ˈɣra.fo.] ~ [ðja.ˈɣra.fo.] 'I delete', there is variation between a full vowel and a glide (see also the discussion in Section 2.4.3). By contrast, traditional orthographic syllabification rules[5] treat such vowel sequences as heterosyllabic, e.g., *μά.τι.α*, *α.λή.θει.α*. Similarly, letter combinations such as *γγ*, which correspond phonemically to a single consonant [g] (e.g., *α.γγα.ρεί.α* [a.ga.ˈɾi.a] 'drudgery, chore'), are treated as single phonemic entities. Traditional orthographic conventions split such double letters into two syllables, e.g., *αγ.γα.ρεί.α*, resulting in an orthographically closed syllable *αγ* [aɣ].

IPA transcription and syllabification were carried out concurrently, as both processes are essential for the accurate identification of syllable boundaries and stress placement. Both processes employed a custom R script (Version 4.4.0; R Core Team, 2024) to generate phonetic representations following standard transcription conventions for Modern Greek (Arvaniti, 1999; Revithiadou, 2021). The script implements phonological rules including voicing

[5] We thank an anonymous reviewer for pointing out that these challenges regarding syllable division rules are not addressed even in the most recent pedagogical grammar used in the final two grades of primary school.

assimilation, glide strengthening, vowel deletion, etc., and applies phonological syllabification based on the sonority sequencing principle (Kappa, 2002).

Table 2 shows examples of words transcribed and syllabified in IPA.

**Table 2**

*Phonological annotation: IPA transcription and syllabification*

| Word | Orthographic Syllabification | IPA Syllabification | Gloss |
|---|---|---|---|
| *αγαπητός* | α.γα.πη.τός | [a.ɣa.pi.ˈtos] | 'lovable' |
| *μαρμάρινος* | μαρ.μά.ρι.νος | [mar.ˈma.ɾi.nos] | 'of marble' |
| *αλήθεια* | α.λή.θεια | [a.ˈli.θça] | 'truth' |

*Note.* ˈ' indicates stress on the following syllable, '.' indicates syllable boundary.

### 2.4.2 Word-level syllable templates

Word-level syllable templates were generated using a custom R script that translated IPA symbols and syllable breaks into syllabic templates. We employed an established notation system where V represents a vowel, C represents a consonant, diphthongs are marked as VJ or JV depending on their position (where J represents a glide), and in coda position, L and N designate liquids and nasals, respectively.

Syllabified word templates provide an abstract representation of syllable structure that can be used for phonotactic analysis or stimulus selection in psycholinguistic experiments. Table 3 shows the syllable template annotation for the same three words.

**Table 3**

*Phonological annotation: Word-level syllable templates*

| Word | CV Template | Gloss |
|---|---|---|

| | | |
|---|---|---|
| *αγαπητός* | V.CV.CV.ˈCVC. | ‘lovable’ |
| *μαρμάρινος* | CVL.CV.CV.ˈCVC. | ‘of marble’ |
| *αλήθεια* | V.ˈCV.CCV. | ‘truth’ |
| *αηδόνι* | VJ.ˈCV.CV. | ‘nightingale’ |

### 2.4.3 Syllable count and stress annotation

Syllable count and stress annotation required both orthographic syllabification and IPA transcription as prerequisites. An R script counted syllables and identified stress position by locating the primary stress mark (the IPA primary stress marker ‘ˈ’ in phonemic transcription). Words were classified as U (ultimate stress, on the final syllable), PU (penultimate stress, on the second-to-last syllable), APU (antepenultimate stress, on the third-to-last syllable), double-stressed (APU and U), or unstressed.

Several issues emerged during stress annotation that required special handling. First, monosyllabic content words such as *γη* [ˈʝi] ‘earth’ and *ζω* [ˈzo] ‘live’ bear phonological stress but lack orthographic accent marks in Greek orthography according to standard spelling conventions, leading to their initial misidentification as stressless by automated tools. Second, words written entirely in capital letters (e.g., *ΑΓΑΛΜΑ* [ˈaɣalma] ‘statue’) lack accent marks and were therefore incorrectly categorized as stressless. Third, words bearing clitic stress patterns, such as *άλογό* [ˈaloˈɣo] ‘horse’, carry double accent marks and were initially misidentified as having U stress. In fact, the correct pattern is APU stress, with the U stress representing clitic stress. Such items are indicated as “Double Stress” in the database.

For monosyllabic words, the script verified PoS category and assigned stress to content words while marking function words as unstressed. Capitalized words were compared manually with their lowercase versions to recover stress information. In cases of capitalization-induced ambiguity (e.g., *ΑΛΛΑ* [ˈala] ‘other (plural neuter) vs. [aˈla] ‘but’), stress was assigned

according to the stress pattern of the more frequent lowercase form, as established by frequency counts in the database.

A further complication in syllable counting and stress assignment arose from variation in the phonetic rendering of certain Greek words. The same word may have different pronunciations due to free variation or speaker-dependent factors. Variation occurs particularly in hiatus environments involving sequences such as *ια, ιά, ιο, ιό, ιού*, etc., where vowel sequences can be realized either as separate syllables or as diphthongs/glide-vowel sequences (see discussion in Section 2.4.1 and references therein). This variation affects syllable count, syllabification patterns, and stress classification. For instance, a word that is APU-stressed in a four-syllable pronunciation becomes PU-stressed in a three-syllable pronunciation; e.g., *α.κρί.βεια* [a.ˈkri.vja] 'precision, high cost' has three syllables and is PU-stressed, whereas *α.κρί.βει.α* [a.ˈkri.vi.a] has four syllables and is APU-stressed. Such words were annotated for both stress patterns. The reported transcriptions for such forms were determined through team discussion and consensus. Approximately 3.2% of word types received dual annotations for pronunciation variation, predominantly involving words with vowel sequences in hiatus positions.

Table 4 presents the syllable count annotation and stress patterns for representative examples.

**Table 4**

*Phonological annotation: Number of syllables and position of stress*

| Word | No. of Syllables | Orthographic Syllabification | IPA Syllabification | Stress |
|---|---|---|---|---|
| *αγαπητός* | 4 | α.γα.πη.τός | [a.ɣa.pi.ˈtos] | U |
| *μαρμάρινος* | 4 | μαρ.μά.ρι.νος | [mar.ˈma.ɾi.nos] | APU |

| | | | | |
|---|---|---|---|---|
| *αλήθεια* | 3 | α.λή.θεια | [a.ˈli.θça] | PU |
| *άλογό* | 3 | ά.λο.γό | [ˈa.lo.ˈɣo] | Double Stress |
| *το* | 1 | *το.* | [to] | Unstressed |
| *ακρίβεια* | 3 / 4 | α.κρί.βεια / α.κρί.βει.α | [a.ˈkri.vja] [a.ˈkri.vi.a] | PU/APU |
| *αντρειοσύνη* | 4 / 5 | α.ντρειο.σύ.νη / α.ντρει.ο.σύ.νη | [a.ⁿdrjo.ˈsi.ni] [a.ⁿdri.o.ˈsi.ni] | PU |

APU: Antepenultimate stress, PU: Penultimate stress, U: Ultimate stress

*Glosses*: lovable, of marble, truth, horse, the (singular neuter), precision / high cost, courage/bravery

### 2.5 Morphological feature annotation

Morphological feature annotation in HelexKids 2.0 focused on noun inflection to extract stress frequency patterns across different noun classes. Because nouns exhibit greater stress-pattern complexity than other word classes, and stress distributions differ across inflection classes (see discussion in Section 1), we annotated nouns in greater detail. Specifically, nouns were manually annotated for their inflectional endings and morphological features including case, number, and gender. A random sample of 1,000 words was independently annotated by three researchers, and an inter-annotator agreement was assessed using Krippendorff's $\alpha$ ($\alpha = 0.827$), indicating strong agreement. Disagreements were resolved through discussion among the team's linguists. Most disagreements arose from the omission of some possible morphosyntactic features that a word type could be associated with. After this process, we established a set of guidelines highlighting the errors most likely to occur and shared them with the annotators. Table 5 presents representative examples of the morphological annotation of nouns.

**Table 5**

*Representative examples of morphological annotation for nouns*

| Word | PoS | Ending | Morphological Features | IPA Syllabification | Gloss |
|---|---|---|---|---|---|
| *άβακα* | NOUN | asM | SgGenAcc | [ˈa.va.ka] | 'abacus' |
| *αβγά* | NOUN | oN | PlNomAcc | [a.ˈvɣa] | 'eggs' |
| *αβιταμίνωση* | NOUN | iF | SgNomAcc | [a.vi.ta.ˈmi.no.si] | 'vitamin deficiency ' |
| *άγαλμα* | NOUN | maN | SgNomAcc | [ˈa.ɣal.ma] | 'statue' |

Sg: Singular, Pl: Plural; Gen: Genitive, Acc: Accusative, Nom: Nominative; M: Masculine, N: Neuter, F: Feminine

## 3 The HelexKids 2.0 database: Structure and interface

In this section, we outline the dynamic redesign of the original HelexKids, which was carried out to make the database more functional and extensible.[6]

### 3.1 Database architecture and storage format

HelexKids 2.0 is implemented as a relational MySQL database consisting of nine interconnected tables organized in a modular architecture that separates linguistic annotation, text- and grade-level metadata, and quantitative measures, allowing users to pose queries across various linguistic dimensions. The central *helexkids_words* table contains 67,802 types (representing over 1.3 million tokens) with comprehensive linguistic annotation. The architecture supports both word-level and database-level analyses through interconnected

[6] For information on HelexKids 2.0 integrity and maintenance, see Appendix A. A complete database schema, field specifications, and technical implementation details are provided in Appendix B.

tables storing pre-computed frequency metrics (including raw frequency, frequency per million, dispersion, and contextual diversity) and book-specific word occurrences. Metadata tables organize the corpus hierarchically by book type, curriculum unit, textbook instruction period, and grade level, while stored query templates enable efficient retrieval of word lists based on these parameters.

This relational structure allows researchers to extract word lists filtered by multiple simultaneous criteria (e.g., syllable count, stress pattern, morphological class, frequency thresholds, grade level) and to examine how vocabulary patterns change across different grade levels. These backend design choices formed the basis for the user-facing tools described in the following section.

### 3.2 Web interface and search capabilities

HelexKids 2.0 is openly accessible to users through a web-based interface hosted on the GRADIENCE project webpage (see fn 4). To accommodate both international and Greek-speaking researchers and educators, the interface is available in English and Greek. It also provides users with direct access to the word frequency database, documentation on updates relative to the original HelexKids, license and citation information, and a bilingual user guide.

The search and filtering interface consists of seven collapsible filter panels that allow users to query the database using multiple criteria simultaneously:

- **Corpus Selection:** Users can select either a single grade (1–6) or a cumulative grade range (1–3 or 4–6), with type and token counts updating dynamically.
- **Part of Speech (PoS):** A scrollable multi-select list allows filtering by one or more grammatical categories. Entries with mixed PoS labels (e.g., ADJ/NOUN) are returned whenever any selected category matches a component of their annotation.

- **Surface Form Search:** Users can retrieve words by internal letter sequence (Word Contains) or orthographic ending (Word Ends With).
- **Syllables and Stress:** A range slider controls syllable count. Stress patterns can be filtered via checkboxes (APU, PU, U, unstressed); a separate dropdown enables filtering by the CV structure of the stressed syllable (e.g., ˈCV, ˈCVC).
- **Frequency and Distribution:** Range sliders control seven standard metrics: Zipf value, frequency per million, log frequency per million, dispersion, and contextual diversity, estimated frequency per million and Standard Frequency Index.
- **Lexical Neighborhood:** Two range sliders implement Levenshtein-based neighborhood density filtering: OLD20 and PLD20.
- **Bigram & Biphone Frequencies**: Four range sliders operating on a logarithmic scale provide filtering by bigram and biphone frequency by type and token count.

Search results are displayed in a sortable table format showing word forms, frequency counts, PoS tags, phonological properties, and (for nouns) morphological features. Results can be sorted by any column to facilitate identification of high-frequency items, rare words, or items with specific phonological properties. The interface supports exporting results in CSV or Excel format with UTF-8 BOM encoding, which ensures proper handling of Greek and IPA characters across different software applications and operating systems.

Results are retrieved from pre-computed frequency metrics stored in the backend to minimize response times and are dynamically filtered according to the criteria specified by the user. Following updates or corrections to the database, the cache files are refreshed to ensure that the metrics remain up to date.

## 4 The HelexKids 2.0 database: Descriptive statistics

In this section, we provide an overview of the descriptive statistics of HelexKids 2.0, including the distribution of PoS and stress patterns, as well as syllable-related measures such as syllable count, average syllable length, and frequent word-level syllable templates. All analyses were conducted in R (Version 4.4.0; R Core Team, 2024).

We begin by examining how lexical coverage develops across the six grade levels represented in the database. Table 6 shows the distribution of the 67,802 word types and also reports token counts as they appear per grade, without deduplication of books shared across grades. We observe that types and tokens increase at every grade, but not at a constant rate. The largest expansion is between Grades 2 and 3 (+87.98% types, +139.91% tokens), due to the increase in the number of school subjects introduced in Grade 3. From this point on, growth rate slows steadily, reaching its lowest point between Grades 5 and 6 (+7.05% types, +14.94% tokens).

**Table 6**

*Word types and tokens per grade (non-cumulative) with grade-to-grade differences (Δ) and relative growth rates (%)*

| Grade | Types (*N*) | Δ Types (*N*) | Δ Types (%) | Tokens (*N*) | Δ Tokens | Δ Tokens (%) |
|---|---|---|---|---|---|---|
| *1* | 9,077 | — | — | 70,345 | — | — |
| *2* | 11,575 | 2,498 | 27.52 | 95,886 | 25,541 | 36.31 |
| *3* | 21,759 | 10,184 | 87.98 | 230,041 | 134,155 | 139.91 |
| *4* | 33,322 | 11,563 | 53.14 | 338,202 | 108,161 | 47.02 |
| *5* | 44,383 | 11,061 | 33.19 | 463,502 | 125,300 | 37.05 |
| *6* | 47,511 | 3,128 | 7.05 | 532,749 | 69,247 | 14.94 |

While Table 6 presents how much material appears at each grade, Table 7 illustrates how much total material a student has encountered by that point in schooling by providing cumulative type and token counts across grade ranges. Interestingly, token counts increase steadily with grade level due to the accumulation of text material over time. However, the degree of increase is not uniform. The smallest raw increase in types occurs early, between Grade 1 and Grades 1–2 (4,530 new types), with every subsequent step adding more. Notably, by Grade 6, where the whole vocabulary and its tokens is available, growth rate is shrunk: 14% for types and 24% for tokens relative to the previous grade step. This is because shared books within grade ranges are counted only at the grade at which they are first introduced (see Section 2). Growth however is still considerable from Grades 1–3 through 1–5. For example, in Grades 1–3, types increase by 49% and tokens by nearly 62% compared to Grades 1–2.

**Table 7**

*Type and token counts across grade ranges (cumulative) with grade-to-grade increments (Δ) and relative growth rates (%)*

| Grade Range | Types (N) | Δ Types | Δ Types (%) | Tokens (N) | Δ Tokens | Δ Tokens (%) |
|---|---|---|---|---|---|---|
| *1* | 9,077 | — | — | 70,345 | — | — |
| *1–2* | 13,607 | 4,530 | 33.3 | 136,880 | 66,535 | 48.62 |
| *1–3* | 26,691 | 13,084 | 49 | 366,921 | 230,041 | 62.7 |
| *1–4* | 40,922 | 14,231 | 34.77 | 637,854 | 270,933 | 42.48 |
| *1–5* | 58,374 | 17,452 | 29.89 | 1,051,783 | 413,929 | 39.35 |
| *1–6* | 67,802 | 9,428 | 13.9 | 1,379,521 | 327,738 | 23.76 |

### 4.1 PoS distribution

The three most frequent categories are nouns (329,644 tokens, 23.90%), verbs (211,743 tokens, 15.35%), and pronouns and articles which form a compound category due to syncretism (PRN/ART: 185,380 tokens, 13.44%). This distribution reflects the discourse characteristics of educational texts, which tend to be noun-heavy and make frequent use of the definite article, a key feature of Greek grammar.

Table 8 presents the frequency distribution of major PoS categories. Content words dominated the lexicon in type counts: nouns (21,606), verbs (20,575), and adjectives (10,452) collectively accounted for nearly 78% of all types. However, these three categories represented just under 44% of total word tokens (nouns: 23.90%; verbs: 15.35%; adjectives: 4.68%). Function words showed the opposite pattern, with minimal vocabulary diversity but high token frequency. For example, articles comprised only 3 types but accounted for 52,102 tokens (3.78%). Similarly, the combined pronoun/article category (PRN/ART) represented just 11 types, yet accounted for 185,380 tokens (13.44% of the database). Prepositions (48 types, 52,435 tokens) and conjunctions (56 types, 80,747 tokens) displayed comparable patterns. (See Appendix C, Table C1 for a complete PoS breakdown.)

**Table 8**

*Distribution of PoS categories across types and tokens in HelexKids 2.0*

| PoS | Types (*N*) | Types (%) | Tokens (*N*) | Tokens (%) |
|---|---|---|---|---|
| *NOUN* | 21,606 | 31.87 | 329,644 | 23.90 |
| *VERB* | 20,575 | 30.35 | 211,743 | 15.35 |
| *PRN/ART* | 11 | 0.02 | 185,380 | 13.44 |
| *CNJ* | 56 | 0.08 | 80,747 | 5.85 |
| *ADJ* | 10,452 | 15.42 | 64,543 | 4.68 |
| *PREP* | 48 | 0.07 | 52,435 | 3.80 |

| | | | | |
|---|---|---|---|---|
| *ART* | 3 | 0.00 | 52,102 | 3.78 |

The stark type–token asymmetry between content and function words reflects a well-established distributional pattern in natural language: nouns, verbs, and adjectives account for most lexical diversity, whereas a small set of high-frequency function words accounts for the majority of total tokens. These differences between types and tokens became more apparent through PoS annotation, transforming the database from a simple word list into a resource for examining how the grammatical patterns of Greek school texts evolve across grades.

### 4.2 Stress pattern distribution

#### 4.2.1 Stress patterns in major parts of speech

As a morphology-dependent system with a three-syllable window restriction, Greek permits stress on the APU, PU, and U syllable. Across all morphosyntactic words in HelexKids 2.0, stress pattern distributions vary systematically. Figure 1 presents the distribution of APU, PU and U stress in verbs, nouns, adjectives, and adverbs, calculated over types and tokens. (Detailed percentage distributions are provided in Appendix C, Tables C2–C3.)

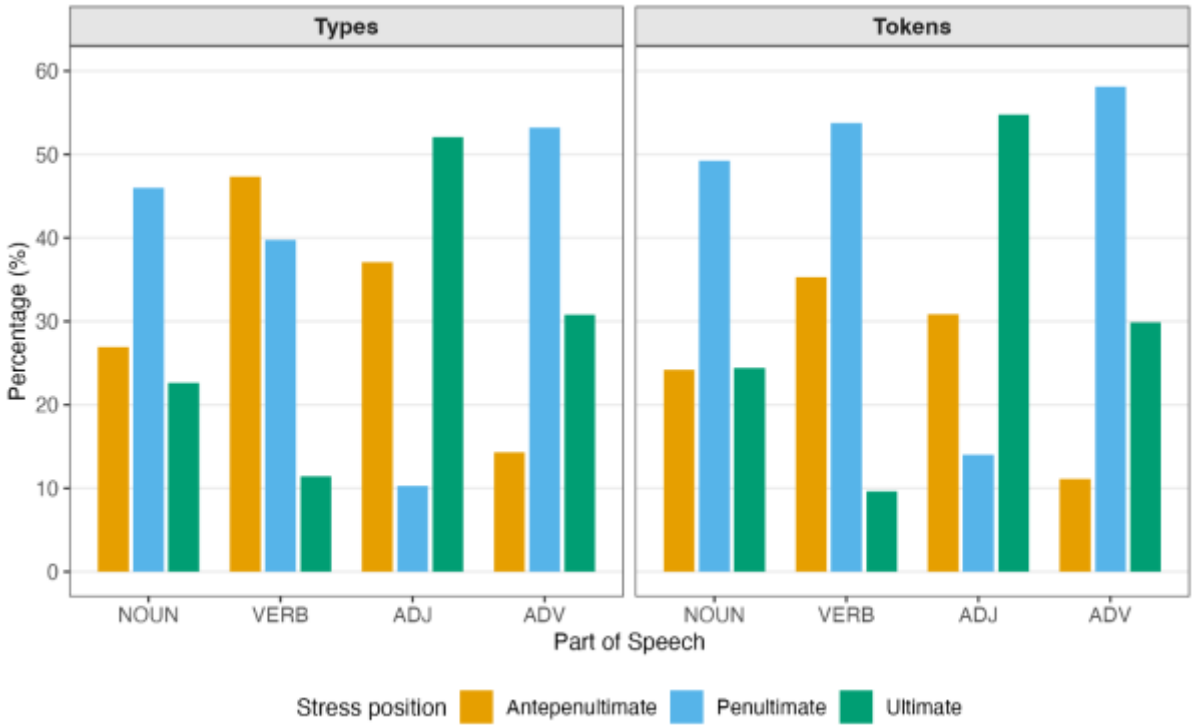


**Figure 1**. Distribution of stress by major PoS in HelexKids 2.0. Left panel shows types, right panel shows tokens. Bars represent the percentage of words with APU (orange), PU (blue), and U (green) stress for nouns, verbs, adjectives, and adverbs.

Nouns exhibit relatively stable distributions across type and token frequencies. They favor PU stress (types: 46.03%; tokens: 49.23%), with APU and U stress occurring at moderate frequencies. Verbs show the highest proportion of APU stress in both types (47.35%) and tokens (35.31%) compared to other morphosyntactic categories. However, types show a closer competition between APU (47.35%) and PU (39.76%) stress than tokens do (PU: 53.75% vs. APU: 35.31%).

Adjectives show a dramatically different pattern: final stress is the predominant stress pattern in both types (52.08%) and tokens (54.81%), while PU stress is relatively low (types: 10.26%; tokens: 14.01%). On the other hand, adverbs show the strongest preference for PU stress (types: 53.22%; tokens: 58.10%); U stress occurs at elevated frequencies compared to

other categories (types: 30.81%; tokens: 29.87%), while APU stress occurs at rather low frequencies (types: 14.29%; tokens: 11.10%).

Such cross-category differences indicate that word class plays a significant role in Greek stress patterns, with each major word category showing distinct distributional properties. This, in turn, underscores the importance of HelexKids 2.0 as a large, developmentally stratified database that enables the systematic investigation of phonology–morphology interactions across word categories and frequency profiles.

### 4.2.2 Stress patterns in nouns and verbs across grade ranges

The grade-level organization of HelexKids 2.0 allows examination of how children's exposure to different stress patterns changes across elementary education. Researchers can investigate stress at individual grade levels (e.g., vocabulary specific to Grade 3 or Grade 5) or at cumulative grade ranges (e.g., all vocabulary encountered through Grade 3 or through Grade 5).

Figure 2 illustrates how stress patterns are distributed across the noun and verb vocabulary that accumulates from Grade 1 through Grade 6. It offers a picture of the specific grammatical categories and their associated stress patterns that students have been exposed to by the end of a given grade. A cumulative perspective better reflects children's vocabulary exposure across schooling. Tracking how stress is distributed within and across grammatical categories, and how this distribution shifts as cumulative vocabulary changes, may also reveal when certain stress patterns become more or less prominent in children's input. (Detailed percentage breakdowns are provided in Appendix C, Tables C4–C7.)

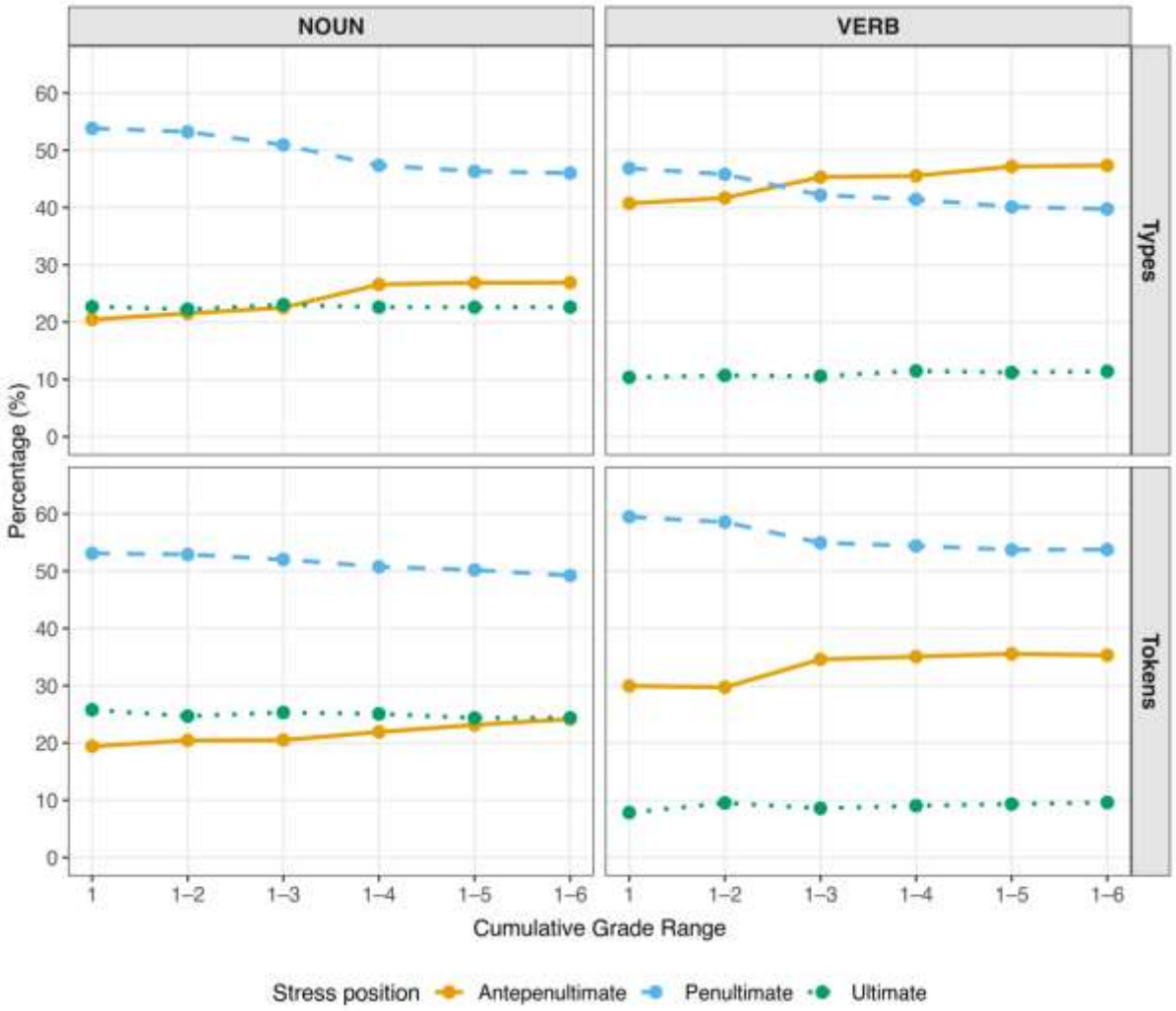


**Figure 2**. Distribution of stress positions in nouns and verbs across cumulative grade ranges in HelexKids 2.0. The top row displays type counts; the bottom row displays token counts. Lines represent APU (solid), PU (dashed), and U (dotted) stress for nouns (left panels) and verbs (right panels) across cumulative grade ranges (Grade 1 to 1–6).

The data reveal distinct patterns for nouns versus verbs that remain relatively stable as vocabulary accumulates across grade levels. For nouns, PU stress consistently prevails across all grade ranges, accounting for 46–54% of types (top left, Figure 2) and 49–53% of tokens (bottom left, Figure 2). APU and U stress show broadly comparable proportions in the type analysis, though they diverge slightly: APU ranges from 20–27%, whereas U remains

remarkably stable at around 23% throughout. In the token-based analysis, U stress is more prominent than APU at Grade 1 (25.78% vs. 19.41%), but the gap closes gradually across grade levels (24.39% vs. 24.17% by Grades 1–6). Notably, both the type and token distributions show a gradual shift across grades. More specifically, from Grade 1 to Grades 1–6, APU stress increases from 20.46% to 26.91% in types and from 19.41% to 24.17% in tokens, while PU stress decreases correspondingly from 53.85% to 46.03% in types and from 53.10% to 49.23% in tokens. U stress remains virtually unchanged across all grade ranges (22.71% to 22.62% in types; 25.78% to 24.39% in tokens). Since the rise in APU (~6–7 percentage points in types, ~5 in tokens) closely mirrors the decline in PU (~8 percentage points in types, ~4 in tokens), the gradual shift in the noun distribution reflects a redistribution specifically between these two stress patterns as the noun vocabulary expands, with U stress not benefiting from this change.

Verbs display a notably different distribution and a more pronounced type-token divergence. In the type analysis, APU and PU stress occur at broadly comparable frequencies, with their relative prominence reversing across grades: APU increases from 40.73% at Grade 1 to 47.35% at Grades 1–6, while PU decreases from 46.86% to 39.76% (top right, Figure 2). The token-based distribution tells a different story: PU-stressed verbs are substantially more frequent in the textbooks throughout (54–59% of tokens), while APU accounts for only 30–36% of tokens (bottom right, Figure 2). This gap – around 12 percentage points between APU types (47.35%) and APU tokens (35.31%) at Grades 1–6, and around 14 percentage points between PU types (39.76%) and PU tokens (53.75%) – indicates that high-frequency verbal forms in children's textbooks are predominantly PU-stressed, despite APU verbal forms being more numerous in the cumulative vocabulary by the later grades. U stress remains low in verbs, representing 10–11% of types and 8–10% of tokens.

Taken together, the systematic type-token divergence observed for verbs but not nouns reveals a fundamental difference between the two categories: the most frequent verbal forms

in children's textbooks are predominantly PU-stressed, whereas for nouns, type and token distributions do not show substantial differences.

### 4.3 Syllable count, average syllable length and most frequent word-level syllable templates across grades

Analysis of HelexKids 2.0 reveals clear grade-level differences in syllable count distribution, average syllable length, and the prevalence of specific syllable templates at the word level. Figure 3 displays the percentage of 1-syllable, 2-syllable, 3-syllable, and 4-syllable words (types) across grade levels. As grade level increases, the proportion of disyllabic words decreases (from 23.1% in Grade 1 to 13.7% in Grade 6). This shift pertains to the relative distribution of word lengths, not the raw frequency with which shorter words appear. Although the proportion of disyllabic words decreases across grades, their raw number increases as the overall database grows, rising from 2,101 words in Grade 1 (out of 9,077 total words) to 6,491 in Grade 6 (out of 47,511 total words). Thus, shorter words do not become less common in absolute terms; instead, they account for a smaller share of a vocabulary that expands with longer words. At the same time, longer words (4+ syllables) increase both in proportion (from 35.1% in Grade 1 to 53.4% in Grade 6) and in absolute count (from 3,183 to 25,356). Three-syllable words show a slight decrease across grades (from 37.5% in Grade 1 to 28.6% in Grade 6), even though their absolute count increases substantially (from 3,403 to 13,604). Words with variable syllable counts (e.g., two or three syllables (2/3)) represent a small proportion of the data (2.2–3.4% across grades). In general, at higher grades, children encounter progressively longer words.

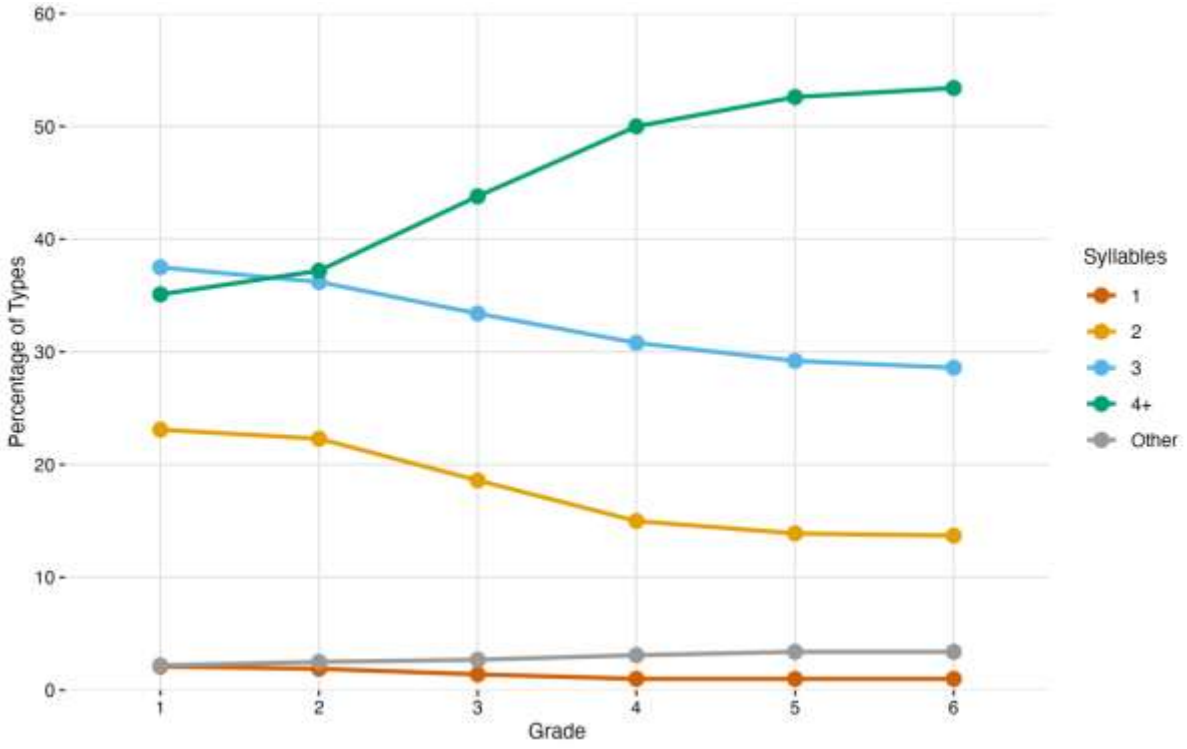


**Figure 3**. Type counts by word-level syllable length across grades 1–6. Longer (4+ syllable) words increase steadily, while 2- and 3-syllable words decline; 1-syllable and "Other" words remain consistently low.

*Note*. "Other" includes words with variable syllable counts (e.g., 2/3).

Additionally, average syllable length increases across grade levels, from 3.25 syllables in Grade 1 (*SD* = 1.11, *n* = 9,077 words) to 3.75 syllables in Grade 6 (*SD* = 1.20, *n* = 47,511 words). Detailed descriptive statistics are reported in Table 9.

**Table 9**

*Mean syllable length of types by grade level, with standard deviations, total word counts, and the proportion of ambiguously coded syllables*

| Grade | Avg Length | SD | Grade Types (*N*) | Ambiguous Syllables (*N*)* | Ambiguous Syllables (%) |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| *1* | 3.249 | 1.11 | 9,077 | 190 | 2.1 |
| *2* | 3.304 | 1.13 | 11,575 | 278 | 2.4 |
| *3* | 3.472 | 1.14 | 21,759 | 587 | 2.7 |
| *4* | 3.661 | 1.18 | 33,322 | 1,036 | 3.1 |
| *5* | 3.729 | 1.20 | 44,383 | 1,495 | 3.4 |
| *6* | 3.754 | 1.20 | 47,511 | 1,597 | 3.4 |

**Note*. Ambiguous cases (e.g., 2/3 syllables) were coded as the midpoint.

Table 10 illustrates that CV.ˈCV.CV. is the most common word template, ranking first consistently across all individual grades. It represents PU-stressed trisyllabic words consisting solely of open syllables.

**Table 10**

*CV.ˈCV.CV. as the most frequent syllable template across all grade levels: type counts and percentages*

| **Grade** | **Template Types (*N*)** | **Template Types (%)** | **Grade Types (*N*)** |
|---|---|---|---|
| *1* | 441 | 4.86 | 9,077 |
| *2* | 481 | 4.16 | 11,575 |
| *3* | 666 | 3.06 | 21,759 |
| *4* | 846 | 2.54 | 33,322 |
| *5* | 1,011 | 2.28 | 44,383 |
| *6* | 1,069 | 2.25 | 47,511 |

*Note*. Percentage reflects the proportion of word types in each grade displaying this template.

The second and third most frequent word templates are ˈCV.CV.CV. (1.78–2.91%) and ˈCV.CV. (1.65–3.9%), with APU and PU stress, respectively. Templates with consonant clusters (e.g., ˈCCV.CV., 0.96–1.83%) are less frequent than those with simple CV structure. (See Appendix C, TableC8 for the top ten templates per grade by types.)

Analysis of word-level syllable templates weighted by token frequency per grade (with no deduplication of shared books) showed that monosyllabic templates were most frequent across all grades. The CV. template ranked first in every grade, accounting approximately for 20–23% of word tokens. Combined, the monosyllabic structures represented in the top 10 (CV., V., CVC., CVN., CCV.) constituted approximately 30–35% of tokens per grade. Among disyllabic templates, ˈCV.CV. was the most frequent (3.38–4.77%), followed by ˈV.CV. (3.35–4.05%) and CV.ˈCV. (2.04–2.93%). The CV.ˈCV.CV. template, which ranked among the most frequent in the type-based analysis, appeared in the token top 10 only for Grades 1 to 4, ranking 7th–10th (1.59–2.47%), and dropped out of the top 10 from Grade 5 onward. (See Appendix C, Table C9, for the top ten templates per grade by tokens.)

## 5 Discussion

### 5.1 Contributions to research on Greek morphophonology

The extended morphophonological annotation in HelexKids 2.0 significantly enhances the resource. For instance, the distributional data on stress patterns across different PoS and inflection classes offer empirical backing for theories of stress assignment in Greek (Malikouti-Drachman & Drachman, 1989; Revithiadou, 1999). Specifically, the predominance of APU stress in specific noun classes (e.g., -*ma*$_N$), as opposed to the preference for PU stress in others (e.g., -*is*$_M$), supports accounts that differentiate between morpholexically conditioned stress and the phonologically determined default (Revithiadou et al., 2026b). On the other hand, the association of verb stress predominantly with PU stress, as opposed to the more even spread of

stress patterns in nouns, supports proposals that nouns are prosodically more complex than verbs (e.g., Smith, 1997, 1999).

Additionally, the morphophonological annotation of lexical databases opens new possibilities for systematically exploring how phonological and morphological properties of words interact. Researchers can test, for example, predictions about systematicity and productivity by examining stress patterns or other phonological properties (e.g., phonotactics, syllable structure) within specific morphosyntactic categories (see, e.g., Dabouis & Fournier, 2023; Storme, 2024). The results of such investigations can in turn refine or challenge existing accounts on the morphophonological characteristics of words, thus contributing to broader debates on lexical representation and grammatical architecture.

Furthermore, HelexKids 2.0 recognizes the gradient and variable nature of phonological outputs by offering dual annotations for words with variable pronunciations, without sacrificing the precision necessary for quantitative analysis. This approach acknowledges the ongoing debates about phonological variation in hiatus environments within Greek (Apostolopoulou, 2018; Markopoulos & Apostolopoulou, 2020; Rytting, 2005), and serves as a point of reference for similar annotation projects.

### 5.2 Linguistic research and educational applications of HelexKids 2.0

From a linguistic perspective, the grade-level organization of HelexKids 2.0 allows researchers to examine how children's exposure to phonological and morphological patterns changes across elementary education. Morphological diversity increases across grades, with longer words appearing in later grades. As a result, morphophonological patterns may be acquired gradually from simpler to more complex forms, through exposure to progressively more advanced vocabulary. Similar findings have been reported among English-speaking elementary school learners ($N$ = 1,052): spelling accuracy of morphemes develops sharply as children

move from Grade 2 to Grade 3, due to the more structured teaching of morphemes from Grade 3 onward (Niolaki et al., 2025).

The database also offers researchers the opportunity to investigate frequency effects in morphophonological acquisition and processing. Since both type and token frequencies are available for all annotated categories, researchers can draw on different frequency measures depending on their research question. For instance, they can examine whether high-frequency stress patterns, word categories or inflection classes are acquired earlier by children (see, e.g., Ambridge et al., 2015; Bybee, 2001, 2010; Tomasello, 2003) or have a processing advantage compared to low frequency structures (see, e.g., Brysbaert et al., 2018). Distributional input statistics (e.g., bigram/trigram frequencies) could also be used to test their role in identifying morphological sub-word units (Rastle & Davis, 2003).

Beyond research applications, HelexKids 2.0 also has important pedagogical value. Unlike speech-based databases, a written corpus such as HelexKids 2.0 reflects the type of vocabulary children encounter in school texts, namely, words that are more academically oriented or specialized depending on the topic of instruction. It is therefore especially useful for reading and spelling support, where exposure to written language plays a central role. For early primary grades, repeated exposure to high-frequency words and recurring orthographic patterns supports direct word recognition in more experienced readers and orthographic learning (Rau et al., 2014; van Viersen et al., 2022). Word frequency also predicts spelling accuracy across primary grades, with its contribution increasing with literacy experience (e.g., Lété et al., 2008). Instruction can gradually include less frequent vocabulary, thus covering more morphologically complex structures such as derived words, whose morphological structure has been shown to facilitate word recognition in children (e.g., Quémart et al., 2012).

The stress pattern annotations are particularly useful for Greek educators, as stress position is a well-documented difficulty for Greek children in both spelling and reading

(Protopapas et al., 2013). Teachers can use HelexKids 2.0 to generate word lists with specific stress patterns for targeted stress instruction, starting from high-frequency word groups with less variable stress patterns before moving to less common ones with more unpredictable stress.

Finally, the PoS and morphological feature annotations also support vocabulary instruction. Teachers can use them to find grade-appropriate examples. Frequency information can again guide this process, given that children are more likely to effectively learn words they encounter repeatedly in their school readings (Nagy & Anderson, 1984).

### 5.3 Comparison with other child-specific databases

HelexKids 2.0 joins a growing collection of child-specific word frequency databases across languages, including *MANULEX* for French (Lété et al., 2004), *childLex* for German (Schroeder et al., 2015), *ESCOLEX* for Portuguese (Soares et al., 2014), and *CYP-LEX* for English (Korochkina et al., 2024), among others. These databases vary in their depth of linguistic annotation: ESCOLEX, childLex, and CYP-LEX provide lemmatization, part-of-speech information, and word length. ESCOLEX additionally includes syllable count and syllable structure information but does not provide phonetic transcriptions.

Complementing this body of work, HelexKids 2.0 extends the annotation scope by including phonological information, serving as a basis for similar annotation practices in other languages. Cross-linguistic comparisons of stress pattern distributions in child lexical databases could illuminate universal and language-specific aspects of children's early vocabulary. Such comparisons have implications for research on how children acquire the phonological and morphological properties of their lexicon. Finally, richly annotated databases are particularly valuable for stimulus selection in experimental designs, as they allow researchers to target or control specific stress patterns (see, for example, Niolaki et al., 2020; Terzopoulos et al., 2020).

### 5.4 Limitations and future directions

Several limitations of the current version of HelexKids 2.0 should be acknowledged. Morphological annotation currently covers only nominal inflection. Future versions should include morphological annotation of the remaining categories to provide a complete morphological profile of all lexical items. Greek verbal morphology shows interesting interactions between stem allomorphy, tense/aspect marking, and stress position (Ralli, 2005; Spyropoulos, 2023), which calls for more detailed verb annotation in the future.

Furthermore, the database is based exclusively on textbook materials. Because textbook content and vocabulary are shaped by curricular, pedagogical, and practical constraints rather than children's productive or receptive vocabulary, the database may not fully capture their actual linguistic exposure. In addition, the materials cover a specific time period (2007–2013), while new textbooks are scheduled for introduction from 2027 onward. As curricula evolve and new digital educational materials have already been approved for use in schools, future versions should consider incorporating more recent educational materials and, of course, materials from children's literature. The expandable architecture of HelexKids 2.0 facilitates future updates, and the annotation pipeline described in this article can be readily extended to new materials with minimal modification.

Finally, the current database focuses on word-level properties and does not include information on syntactic contexts or collocations because the source database consisted of isolated lexical items without sentential context. Context-sensitive annotation and word co-occurrence statistics for exploring the relation between lexical and morphosyntactic patterns remain an important goal for future versions of the database, particularly in light of the new textbooks available from 2027 onwards. Future improvements could also include the addition of lemmatization, which would allow users to access lemma-level frequency data and identify inflectional and derivational families. Such information would greatly enhance the analytical

potential of the database, allowing researchers to move beyond individual word forms and explore vocabulary structure at a deeper morphological level.

## 6 Conclusion

HelexKids 2.0 is a freely accessible annotated resource covering many linguistic and frequency-based dimensions. The sophisticated annotation methodology, which combines automated approaches with extensive manual verification, offers high-quality data for PoS categories, phonological properties, and, in the current version, morphological features for nouns. The database is designed for both research and education. It supports research in developmental psycholinguistics, phonology, morphology, and computational modeling of Greek, while also serving as a useful tool for teachers and curriculum developers.

By making HelexKids 2.0 openly accessible and by continually improving and extending it, we hope it will become a valuable resource for supporting empirical research on the Greek language, children's language development, and evidence-based language instruction.

**Acknowledgements**

We thank the remaining researchers of the original HelexKids team (Lynne Duncan, Mark Wilson, and Jackie Masterson) for providing the foundation for this expanded database. Our sincere thanks also go to the anonymous reviewers and the Associate Editor for their valuable comments and suggestions, which have greatly improved this manuscript. We also wish to thank Panos Argyrakis, Eirini Apostolopoulou, Vassiliki Apostolouda, Michalis Kanetidis, Maria Soukalopoulou, and Maria Tsouchnika for their contribution to HelexKids 2.0, and Vassilios Spyropoulos for expert advice on morphophonological annotation questions. Giorgos Markopoulos was affiliated with the University of the Aegean during the GRADIENCE project (1/1/2024–31/12/2025), within the framework of which this research was conducted. Since 25 February 2026, he has been affiliated with the School of Philology, University of Ioannina. Claude (Anthropic, claude-sonnet-4-6) was used for proofreading and minor phrasing suggestions, all of which were reviewed and approved by the authors. All remaining errors are our own.

**Declarations**

**Funding**

The research leading to the development of HelexKids 2.0 was conducted as part of the project *GRADIENCE – Modeling the limits of grammar: Integrating lexical frequency in a Gradient Harmonic model of lexical stress; Evidence from young and adult Greek speakers' grammars* (Project H.F.R.I. ID: 015053, A.U.Th. R.C. ID: 76809; https://gradience.lit.auth.gr; P.I. Anthi Revithiadou). This project was carried out within the framework of the National Recovery and Resilience Plan Greece 2.0, funded by the European Union – NextGenerationEU (Implementation body: H.F.R.I.; Beneficiary: A.U.Th.).

**Conflicts of interest/Competing interests**

The authors have no relevant financial or non-financial interests to disclose. The authors have no competing interests to declare that are relevant to the content of this article.

**Ethics approval**

Not applicable.

**Consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Availability of data and materials**

The database and all relevant materials are freely accessible at https://gradience.lit.auth.gr/index.php/helexkids-2-0/ and https://gradience.lit.auth.gr/index.php/helexkids_2-0/.

**Code availability**

The codes used for the development of HelexKids 2.0 are freely accessible at: https://gradience.lit.auth.gr/index.php/helexkids_2-0/.

**Open Practices**

Materials and analysis codes are available at https://gradience.lit.auth.gr/index.php/helexkids_2-0/. Preregistration is not applicable.

**References**

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17*(9), 814–823. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Alderete, J. D. (1999). *Morphologically governed accent in Optimality Theory* (Doctoral dissertation, University of Massachusetts Amherst). https://doi.org/10.7282/T3J965B0

Alderete, J. D. (2001). *Morphologically governed accent in Optimality Theory*. Routledge. https://doi.org/10.4324/9781315054834

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*(2), 239–273. https://doi.org/10.1017/S030500091400049X

Anastassiadis-Symeonidis, A. (2012). Το νεοελληνικό κλιτικό σύστημα των ουσιαστικών και οι τάσεις του [The Modern Greek noun inflectional system and its tendencies]. In Z. Gavriilidou, A. Efthymiou, E. Thomadaki, & P. Kambakis-Vougiouklis (Eds.), *Selected papers of the 10th International Conference of Greek Linguistics* (pp. 23–40). Democritus University of Thrace.

Anastassiadis-Symeonidis, A., & Fliatouras, A. (2004). Το λόγιο και το λαϊκό στην ελληνική γλώσσα: Ορισμός και ταξινόμηση [The learned and the colloquial in the Greek language: Definition and classification]. In G. Katsimali, A. Kalokairinos, E. Anagnostopoulou, & I. Kappa (Eds.), *Proceedings of the 6th International Conference on Greek Linguistics*. University of Crete.

Anastassiadis-Symeonidis, A., & Fliatouras, A. (2018). Από το λόγιο επίπεδο της Νέας Ελληνικής στην Αρχαία Ελληνική: Εφαρμογή και διδακτικές προοπτικές [From the learned register of Modern Greek to Ancient Greek: Research proposals and educational perspectives]. *Studies in Greek Linguistics, 38*, 37–50.

Apostolopoulou, E. (2018). *The phonological representation of the glide in Standard Modern Greek: A Gradient Harmonic Grammar approach* [Master's thesis, Aristotle University of Thessaloniki]. https://doi.org/10.26262/heal.auth.ir.301177

Apostolouda, V. (2018). *Πειραματικές διερευνήσεις στον τονισμό της ελληνικής* [Experimental investigations on Greek stress] [Doctoral dissertation, Aristotle University of Thessaloniki]. https://doi.org/10.12681/eadd/44406

Apoussidou, D. (2007). *The learnability of metrical phonology* [Doctoral dissertation, University of Amsterdam]. LOT. https://hdl.handle.net/11245/1.261837

Arvaniti, A. (1999). Standard Modern Greek. *Journal of the International Phonetic Association, 29*(2), 167–172. https://doi.org/10.1017/S0025100300006538

Baltazani, M., & Topintzi, N. (2012). On some phonetic and phonological properties of the Greek glide. In Z. Gavriilidou, A. Efthymiou, E. Thomadaki, & P. Kambakis-Vougiouklis (Eds.), *Proceedings of the 10th International Conference on Greek Linguistics* (pp. 153–166). Democritus University of Thrace.

Baltazani, M., Kainada, E., Revithiadou, A., & Topintzi, N. (2016). Vocoid-driven processes: Palatalization and glide hardening in Greek and its dialects. *Glossa: A Journal of General Linguistics, 1*(1), 23. https://doi.org/10.5334/gjgl.108

Beckman, M. E., & Edwards, J. (2000). The ontogeny of phonological categories and the primacy of lexical learning in linguistic development. *Child Development, 71*(1), 240–249. https://doi.org/10.1111/1467-8624.00139

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science, 27*(1), 45–50. https://doi.org/10.1177/0963721417727521

Burzio, L., & Tantalou, N. (2007). Modern Greek accent and faithfulness constraints in OT. *Lingua, 117*(6), 1080–1124. https://doi.org/10.1016/j.lingua.2006.05.008

Bybee, J. (2001). *Phonology and language use*. Cambridge University Press.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.

Charalambakis, C. (Ed.). (2023). *Practical dictionary of Modern Greek* (2nd ed.). Academy of Athens. https://christikolexiko.academyofathens.gr

Dabouis, Q., & Fournier, J.-M. (2023). The stress patterns of English verbs: Syllable weight and morphology. In E. Yamada, A. Przewozny, J.-M. Fournier, & N. Ballier (Eds.), *New perspectives on English word stress* (pp. 154–191). Edinburgh University Press. https://doi.org/10.1515/9781399519953-011

Dimitropoulou, M., Duñabeitia, J. A., Avilés, A., Corral, J., & Carreiras, M. (2010). Subtitle-based word frequencies as the best estimate of reading behavior: The case of Greek. *Frontiers in Psychology, 1*, Article 218. https://doi.org/10.3389/fpsyg.2010.00218

Drachman, G., & Malikouti-Drachman, A. (1999). Greek word accent. In H. van der Hulst (Ed.), *Word prosodic systems in the languages of Europe* (pp. 897–946). Mouton de Gruyter. https://doi.org/10.1515/9783110197082.2.897

Fliatouras, A., & Anastassiadis-Symeonidis, A. (Eds.). (2019). *Το λόγιο επίπεδο στη Σύγχρονη Νέα Ελληνική: Θεωρία, ιστορία, εφαρμογή* [The learned register in Contemporary Modern Greek: Theory, history, application]. Patakis.

Holton, D., Mackridge, P., Philippaki-Warburton, I., & Spyropoulos, V. (2012). *Greek: A comprehensive grammar of the modern language* (2nd ed.). Routledge.

Idsardi, W. J. (1992). *The computation of prosody* [Doctoral dissertation, Massachusetts Institute of Technology].

Institute for Language and Speech Processing - Athena Research Center. (2015). *Hellenic National Corpus*. Version 3.0. [Dataset (Text corpus)]. CLARIN:EL. http://hdl.handle.net/11500/ATHENA-0000-0000-23E2-9

Institute of Modern Greek Studies, Manolis Triantafyllidis Foundation. (1998). *Dictionary of Standard Modern Greek*.

Kahn, D. (1976). *Syllable-based generalizations in English phonology* [Doctoral dissertation, Massachusetts Institute of Technology].

Kappa, I. (2002). On the acquisition of syllabic structure in Greek. *Journal of Greek Linguistics, 3*(1), 1–52. https://doi.org/10.1075/jgl.3.03kap

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods, 42*(3), 643–650. https://doi.org/10.3758/BRM.42.3.643

Kiparsky, P. (1973). The inflectional accent of Indo-European. *Language, 49*(4), 794–849. https://doi.org/10.2307/412064

Kiparsky, P. (1984). Lexical phonology of Sanskrit word accent. In S. D. Joshi (Ed.), *Amṛtadhārā: R. N. Dandekar felicitation volume* (pp. 201–210). Ajanta.

Kiparsky, P., & Halle, M. (1977). Towards a reconstruction of the Indo-European accent. In L. M. Hyman (Ed.), *Studies in stress and accent* (pp. 209–238). University of Southern California.

Korochkina, M., Marelli, M., Brysbaert, M., & Rastle, K. (2024). The children and young people's books lexicon (CYP-LEX): A large-scale lexical database of books read by children and young people in the United Kingdom. *Quarterly Journal of Experimental Psychology, 77*(12), 2418–2438. https://doi.org/10.1177/17470218241229694

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*(3), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

Kyparissiadis, A., van Heuven, W. J. B., Pitchford, N. J., & Ledgeway, T. (2017). GreekLex 2: A comprehensive lexical database with part-of-speech, syllabic, phonological, and stress information. *PLOS ONE, 12*(2), Article e0172493. https://doi.org/10.1371/journal.pone.0172493

Lété, B., Peereman, R., & Fayol, M. (2008). Consistency and word-frequency effects on spelling among first- to fifth-grade French children: A regression-based study. *Journal of Memory and Language, 58*(4), 952–977. https://doi.org/10.1016/j.jml.2008.01.001

Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, & Computers, 36*(1), 156–166. https://doi.org/10.3758/BF03195560

Malikouti-Drachman, A., & Drachman, G. (1989). Stress in Greek. *Studies in Greek Linguistics, 9*, 127–143.

Markopoulos, G., & Apostolopoulou, E. (2020). When grammar gets noisy: A comparative study of register-driven variation. *Studies in Greek Linguistics, 40*, 303–312.

Massaro, D. W. (2015). Two different communication genres and implications for vocabulary development and learning to read. *Journal of Literacy Research, 47*(4), 505–527. https://doi.org/10.1177/1086296X15616120

Melvold, J. L. (1990). *Structure and stress in the phonology of Russian* [Doctoral dissertation, Massachusetts Institute of Technology].

Montag, J. L., & MacDonald, M. C. (2015). Text exposure predicts spoken production of complex sentences in 8- and 12-year-old children and adults. *Journal of Experimental Psychology: General, 144*(2), 447–468. https://doi.org/10.1037/xge0000054

Nagy, W. E., & Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly, 19*(3), 304–330. https://doi.org/10.2307/747823

Niolaki, G., Vousden, J., Terzopoulos, A., Purser, H., Kohnen, S., & Masterson, J. (2025). A qualitative exploration of spelling errors using the Spelling Profile Assessment: A new UK spelling test for primary-age children. *PATOSS Bulletin, Winter*.

Niolaki, G. Z., Terzopoulos, A. R., & Masterson, J. (2020). Intervention for a visual attention span processing deficit in a Greek-speaking child with slow reading speed. *Preschool and Primary Education*, *7*(3). https://doi.org/10.12681/ppej.21181

Petrounias, E. (2002). *Νεοελληνική γραμματική και συγκριτική ανάλυση, Μέρος Α΄: Θεωρία* [Modern Greek grammar and comparative analysis, Part A: Theory] (2nd ed.). Ziti.

Protopapas, A., Fakou, A., Drakopoulou, S., Skaloumbakas, C., & Mouzaki, A. (2013). What do spelling errors tell us? Classification and analysis of errors made by Greek schoolchildren with and without dyslexia. *Reading and Writing, 26*(5), 615–646. https://doi.org/10.1007/s11145-012-9378-3

Protopapas, A., Tzakosta, M., Chalamandaris, A., & Tsiakoulis, P. (2012). IPLR: An online resource for Greek word-level and sublexical information. *Language Resources and Evaluation, 46*(3), 449–459. https://doi.org/10.1007/s10579-010-9130-z

Quémart, P., Casalis, S., & Duncan, L. G. (2012). Exploring the role of bases and suffixes when reading familiar and unfamiliar words: Evidence from French young readers. *Scientific Studies of Reading, 16*(5), 424–442. https://doi.org/10.1080/10888438.2011.584333

Rastle, K., & Davis, M. H. (2003). Reading morphologically complex words: Some thoughts from masked priming. In S. Kinoshita & S. J. Lupker (Eds.), Masked priming: The state of the art (pp. 279–305).

R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.4.2) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Ralli, A. (2000). A feature-based analysis of Greek nominal inflection. *Glossologia, 11–12*, 201–227.

Ralli, A. (2005). *Morphology*. Patakis.

Ralli, A., & Touratzidis, L. (1991). Υπολογιστική επεξεργασία του τονισμού της Νέας Ελληνικής [A computational processing of Modern Greek stress]. *Studies in Greek Linguistics, 11*, 273–292.

Rau, A. K., Moeller, K., & Landerl, K. (2014). The transition from sublexical to lexical processing in a consistent orthography: An eye-tracking study. *Scientific Studies of Reading, 18*(3), 224–233. https://doi.org/10.1080/10888438.2013.857673

Revithiadou, A. (1999). *Headmost accent wins: Head dominance and ideal prosodic form in lexical accent systems* [Doctoral dissertation, Holland Institute of Generative Linguistics, Leiden University]. Holland Academic Graphics. https://doi.org/10.7282/T3MS3RJ1

Revithiadou, A. (2021). *Εισαγωγή στη Φωνολογία* [Introduction to phonology]. Institute of Modern Greek Studies, Manolis Triandafyllidis Foundation.

Revithiadou, A., & Lengeris, A. (2016). One or many? In search of the default stress in Greek. In J. Heinz, R. Goedemans, & H. van der Hulst (Eds.), *Dimensions of phonological stress* (pp. 263–290). Cambridge University Press. https://doi.org/10.1017/9781316212745.010

Revithiadou, A., Markopoulos, G., Kosmidis, K., Mittas, I., Soukalopoulou, M., & Apostolouda, V. (2026a). Grammar over lexical frequency: Morphological conditioning and the phonological default in Greek stress assignment [Manuscript submitted for publication]. Aristotle University of Thessaloniki & University of the Aegean.

Revithiadou, A., Soukalopoulou, M., Apostolouda, V., & Paspali, A. (2026b). Eliciting the grammar of stress: Morphophonological conditioning in Greek by school-age children and adults [Manuscript submitted for publication]. Aristotle University of Thessaloniki.

Rytting, C. A. (2005). An iota of difference: Attitudes to yod in lexical and social contexts. *Journal of Greek Linguistics, 6*, 151–185. https://doi.org/10.1075/jgl.6.08ryt

Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A., & Kliegl, R. (2015). childLex: A lexical database of German read by children. *Behavior Research Methods, 47*(4), 1085–1094. https://doi.org/10.3758/s13428-014-0528-1

Selkirk, E. (1982). The syllable. In H. van der Hulst & N. Smith (Eds.), *The structure of phonological representations* (Vol. 2, pp. 337–383). Foris.

Sims, A. D. (2019). When the default is exceptional: Word stress in Modern Greek nouns. In J. J. Pennington, V. A. Friedman, & L. A. Grenoble (Eds.), *And thus you are everywhere honored: Studies dedicated to Brian D. Joseph* (pp. 321–347). Slavica Publishers.

Smith, J. (1999). Noun faithfulness and accent in Fukuoka Japanese. In S. Bird, A. Carnie, J. Haugen, & P. Norquest (Eds.), *Proceedings of the 18th West Coast Conference on Formal Linguistics* (pp. 519–531). Cascadilla Press.

Smith, J. L. (1997). *Noun faithfulness: On the privileged behavior of nouns in phonology* [Unpublished manuscript]. University of Massachusetts, Amherst / Rutgers Optimality Archive.

Soares, A. P., Medeiros, J. C., Simões, A., Machado, J., Costa, A., Pinheiro, A. P., & Comesaña, M. (2014). ESCOLEX: A grade-level lexical database from European Portuguese elementary to middle school textbooks. *Behavior Research Methods, 46*(1), 240–253. https://doi.org/10.3758/s13428-013-0350-1

Spyropoulos, V. (2023). Το ρηματικό σύστημα της Ελληνικής [The verbal system of Greek]. In D. Papadopoulou & A. Revithiadou (Eds.), *Εισαγωγή στη Μορφολογία* [Introduction to morphology] (pp. 150–198). Institute of Modern Greek Studies (Manolis Triantafyllidis Foundation).

Storkel, H. L. (2004). The emerging lexicon of children with phonological delays: Phonotactic constraints and probability in acquisition. *Journal of Speech, Language, and Hearing Research, 47*(5), 1194–1212. https://doi.org/10.1044/1092-4388(2004/089)

Storme, B. (2024). French liaison is allomorphy, not allophony: Evidence from lexical statistics. *Morphology, 34*, 415–455. https://doi.org/10.1007/s11525-024-09429-8

Suárez, L., Tan, S. H., Yap, M. J., & Goh, W. D. (2011). Observing neighborhood effects without neighbors. *Psychonomic Bulletin & Review, 18*, 605–611. https://doi.org/10.3758/s13423-011-0078-9

Terzopoulos, A. R., Duncan, L. G., Wilson, M. A. J., Niolaki, G. Z., & Masterson, J. (2017). HelexKids: A word frequency database for Greek and Cypriot primary school children. *Behavior Research Methods, 49*(1), 83–96. https://doi.org/10.3758/s13428-015-0698-5

Terzopoulos, A. R., Niolaki, G. Z., & Masterson, J. (2020). Intervention for a lexical reading and spelling difficulty in two Greek-speaking primary-age children.

*Neuropsychological Rehabilitation, 30*(3), 371–392. https://doi.org/10.1080/09602011.2018.1467330

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Topintzi, N., & Baltazani, M. (2016). The third dream: Aspects of the morphophonology of the Greek glide and palatals. *Linguistics Unlimited, 1*, 1–27.

Triantafyllidis, M. (2018 [1941]). *Modern Greek grammar*. Institute of Modern Greek Studies (Manolis Triantafyllidis Foundation).

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology, 67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

van Viersen, S., Protopapas, A., Georgiou, G. K., Parrila, R., Ziaka, L., & de Jong, P. F. (2022). Lexicality effects on orthographic learning in beginning and advanced readers of Dutch: An eye-tracking study. *Quarterly Journal of Experimental Psychology, 75*(6), 1135–1154. https://doi.org/10.1177/17470218211047420

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language, 40*(3), 374–408. https://doi.org/10.1006/jmla.1998.2618

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language, 60*(4), 502–529. https://doi.org/10.1016/j.jml.2009.02.001

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*(5), 971–979. https://doi.org/10.3758/PBR.15.5.971

**List of abbreviations and symbols**

| | |
|---|---|
| Acc | Accusative |
| ADJ | Adjective |
| ADV | Adverb |
| APU | Antepenultimate stress |
| ART | Article |
| C | Consonant |
| CD | Contextual Diversity |
| CNJ | Conjunction |
| D | Dispersion |
| EXCL | Exclamation |
| Gen | Genitive |
| GER | Gerund |
| IPA | International Phonetic Association |
| J | Glide |
| U | Ultimate stress |
| u | Estimated frequency per million |
| V | Vowel |
| L | Liquid consonant |
| N | Nasal consonant |
| NFC | Unicode Normalization Form C (Canonical Decomposition, followed by Canonical Composition) |
| Nom | Nominative |
| Pl | Plural |
| PNOUN | Proper noun |
| PoS | Part of Speech |
| PREP | Preposition |
| PRN | Pronoun |
| PU | Penultimate stress |
| SD | Standard Deviation |
| SFI | Standard Frequency Index |
| Sg | Singular |

# Appendices

## Appendix A. Database integrity and maintenance

To ensure long-term database integrity and usability, we have implemented several safeguards:

- **Version control**: The database is maintained under version control with clear documentation of any corrections or updates. The current version is HelexKids 2.0, 31/12/2025.
- **Error reporting mechanism**: Users can report suspected annotation errors or inconsistencies through the GRADIENCE email address (gradience@lit.auth.gr). Reported issues are reviewed by the research team, and corrections are incorporated into periodic database updates.
- **Archival plan**: The HelexKids 2.0 database is hosted and maintained on the Aristotle University of Thessaloniki (A.U.Th.) website, where it is made publicly available via a stable institutional URL to ensure long-term access and citability.
- **License**: *HelexKids* 2.0 is released under a Creative Commons Attribution–NonCommercial 4.0 International (CC BY-NC 4.0) license, which permits non-commercial use, distribution, and adaptation of the database, provided that appropriate credit is given to the authors and the source. Users are instructed to cite both the present article and the original *HelexKids* publication (Terzopoulos et al., 2017) when using the database.

## Appendix B. The architecture of HelexKids 2.0

The HelexKids 2.0 database architecture is designed to support simple and complex linguistic queries across six grade levels. Figure B1 presents the internal structure of the database, and below we describe each table and its role in the database ecosystem.

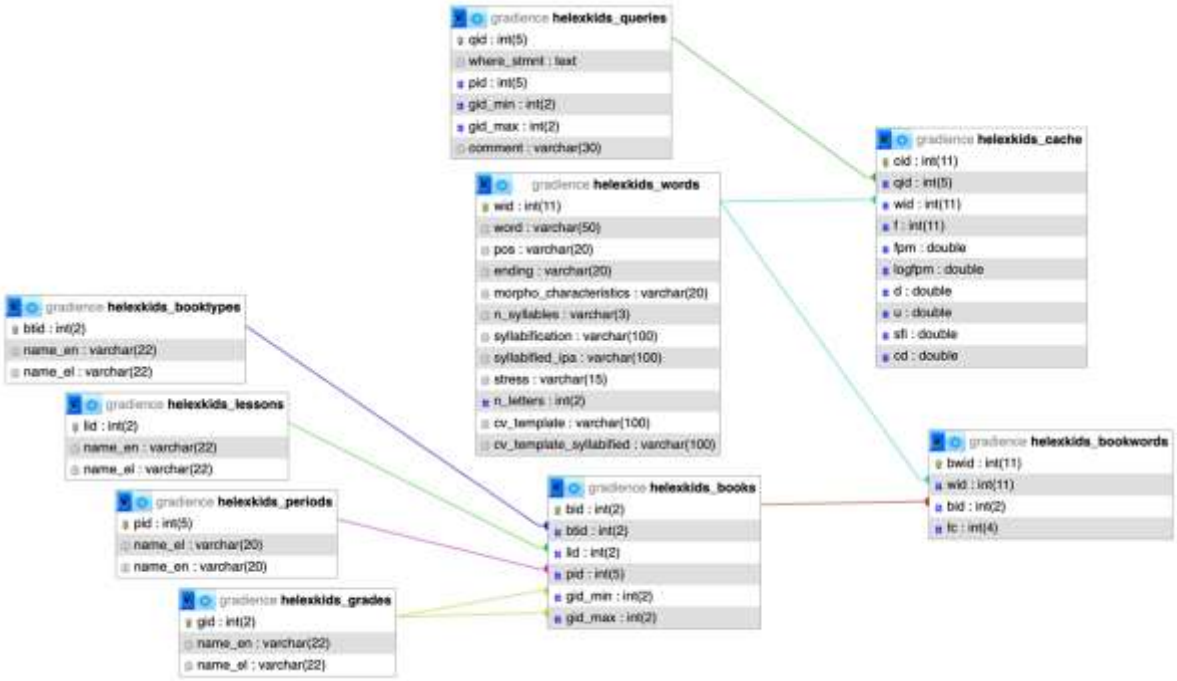


**Figure B1**. Relational schema of HelexKids 2.0, with lexical entries linked to educational textbooks, periods of textbook instruction, grade ranges, morphophonological annotations, and precomputed frequency measures.

*Note.* Connecting lines indicate primary and foreign key relationships between tables.

### Core data tables:

*helexkids_words* (Central table): Contains 67,802 types with linguistic annotation. Each word is identified by a type ID (wid: int(11)) and includes: orthographic form (word: varchar(50)), part of speech (pos: varchar(20)), inflectional ending (ending: varchar(20)), morphological feature string (morpho_characteristics: varchar(20)), syllable count (n_syllables: varchar(3)),

syllabified form (syllabification: varchar(100)), IPA transcription with syllable boundaries (syllabified_ipa: varchar(100)), stress pattern (stress: varchar(15)), letter count (n_letters: int(2)), consonant-vowel template (cv_template: varchar(100)), and syllabified CV template (cv_template_syllabified: varchar(100)). This detailed annotation allows filtering by phonological, morphological, and orthographic criteria.

*helexkids_cache*: Stores pre-computed frequency and distributional statistics for word-query combinations. Each cache entry is identified by a cache ID (cid: int(11)) and links to a query ID (qid: int(5)) and word ID (wid: int(11)), containing: raw frequency count (f: int(11)), frequency per million words (fpm: double), log-transformed frequency per million (logfpm: double), dispersion index measuring spread across texts (d: double), usage coefficient (u: double), standard frequency index (sfi: double), and contextual diversity measuring the number of different contexts (cd: double). Pre-computing these metrics is necessary to improve query performance for large-scale frequency analyses. Periodic updates are performed by the database administrator.

*helexkids_bookwords* (Junction table): Links individual words to specific textbooks. Each entry is identified by a bookword ID (bwid: int(11)) and connects a word ID (wid: int(11)) to a book ID (bid: int(2)), storing the frequency count (fc: int(4)) of that word in that book. This table enables both aggregated database-level frequency calculations and book-specific vocabulary analyses.

**Metadata and classification tables:**

*helexkids_books*: Bibliographic metadata for 76 textbooks and workbooks in the database. Each book is identified by a book ID (bid: int(2)) and is classified by: book type ID (btid: int(2),

linking to helexkids_booktypes), lesson ID (lid: int(2), linking to helexkids_lessons), period ID (pid: int(5), linking to helexkids_periods), and grade range (gid_min, gid_max: int(2), linking to helexkids_grades). This multi-dimensional classification allows for sophisticated filtering by content type.

*helexkids_booktypes*: Lookup table for book type categories (e.g., language and arts textbooks, readers, workbooks) with book type ID (btid: int(2)) and bilingual labels (name_en, name_el: varchar(22)).

*helexkids_lessons*: Lookup table for curriculum units or lesson types, identified by lesson ID (lid: int(2)) with bilingual labels (name_en, name_el: varchar(22)). This structure enables filtering by pedagogical content area.

*helexkids_periods*: Defines the instructional period of the stored textbooks, identified by period ID (pid: int(5)) with bilingual labels (name_el, name_en: varchar(20)).

*helexkids_grades*: Grade level classifications corresponding to the Greek educational system, identified by grade ID (gid: int(2)) with bilingual labels (name_en, name_el: varchar(22)).

**Query management and storage:**

*helexkids_queries*: Stores parameterized query templates for common data extraction patterns. Each query is identified by a query ID (qid: int(5)).

## Appendix C. Extended descriptive statistics

### Table C1

*Frequency distribution of PoS categories across types and tokens in HelexKids 2.0*

| PoS | Types (*N*) | Types (%) | Tokens (*N*) | Tokens (%) |
|---|---|---|---|---|
| *NOUN* | 21,606 | 31.87 | 329,644 | 23.90 |
| *VERB* | 20,575 | 30.35 | 211,743 | 15.35 |
| *PRN/ART* | 11 | 0.02 | 185,380 | 13.44 |
| *CNJ* | 56 | 0.08 | 80,747 | 5.85 |
| *ADJ* | 10,452 | 15.42 | 64,543 | 4.68 |
| *PREP* | 48 | 0.07 | 52,435 | 3.80 |
| *ART* | 3 | 0.00 | 52,102 | 3.78 |
| *PRN/ADJ* | 218 | 0.32 | 40,634 | 2.95 |
| *PREP+ART* | 10 | 0.01 | 39,272 | 2.85 |
| *ADJ/NOUN* | 2,696 | 3.98 | 36,088 | 2.62 |
| *PNOUN* | 4,869 | 7.18 | 34,772 | 2.52 |
| *CNJ/PTCL* | 3 | 0.00 | 32,283 | 2.34 |
| *ADV* | 357 | 0.53 | 30,156 | 2.19 |
| *PRN* | 48 | 0.07 | 27,703 | 2.01 |
| *PREP/PRN* | 1 | 0.00 | 22,780 | 1.65 |
| *PRN/CNJ* | 13 | 0.02 | 22,542 | 1.63 |
| *ADJ/ADV* | 1,103 | 1.63 | 20,875 | 1.51 |
| *NUM/ART/PRN* | 12 | 0.02 | 19,843 | 1.44 |

| | | | | |
|---|---|---|---|---|
| *PTCL* | 16 | 0.02 | 17,780 | 1.29 |
| *NUM* | 281 | 0.41 | 10,840 | 0.79 |
| *ADV/CNJ* | 10 | 0.01 | 8,036 | 0.58 |
| *OTHER* | 1,911 | 2.82 | 6,725 | 0.49 |
| *PCP/ADJ* | 2,120 | 3.13 | 6,216 | 0.45 |
| *VERB/NOUN* | 199 | 0.29 | 4,394 | 0.32 |
| *GER* | 666 | 0.98 | 3,826 | 0.28 |
| *ADJ/ADV/NOUN* | 95 | 0.14 | 3,753 | 0.27 |
| *CNJ/PREP* | 5 | 0.01 | 3,342 | 0.24 |
| *NOUN/PNOUN* | 130 | 0.19 | 1,577 | 0.11 |
| *ADV/DM* | 3 | 0.00 | 1,537 | 0.11 |
| *PREP/ADV* | 3 | 0.00 | 1,306 | 0.09 |
| *ADJ/PNOUN* | 65 | 0.10 | 1,074 | 0.08 |
| *ADV/PTCL/EXCL* | 3 | 0.00 | 957 | 0.07 |
| *PRN/ADJ/ADV* | 2 | 0.00 | 839 | 0.06 |
| *EXCL/DM* | 64 | 0.09 | 759 | 0.06 |
| *PREP/ADV/CNJ* | 2 | 0.00 | 656 | 0.05 |
| *CNJ/DM* | 1 | 0.00 | 606 | 0.04 |
| *ADV/NOUN* | 13 | 0.02 | 454 | 0.03 |
| *ADV/EXCL* | 4 | 0.01 | 391 | 0.03 |
| *PCP/ADJ/NOUN* | 57 | 0.08 | 228 | 0.02 |
| *ADV/PTCL* | 2 | 0.00 | 115 | 0.01 |
| *ADJ/ADV/CNJ* | 1 | 0.00 | 108 | 0.01 |
| *PCP/ADJ/ADV* | 15 | 0.02 | 85 | 0.01 |

| | | | | |
|---|---|---|---|---|
| *NOUN/EXCL/DM* | 5 | 0.01 | 72 | 0.01 |
| *NUM/NOUN* | 5 | 0.01 | 60 | 0.00 |
| *ADJ/NOUN/PNOUN* | 5 | 0.01 | 53 | 0.00 |
| *NUM/ADV* | 2 | 0.00 | 41 | 0.00 |
| *NOUN/GER* | 1 | 0.00 | 26 | 0.00 |
| *ADJ/ADV/PREP* | 2 | 0.00 | 23 | 0.00 |
| *VERB/EXCL/DM* | 2 | 0.00 | 22 | 0.00 |
| *PRN/ADJ/NOUN* | 2 | 0.00 | 20 | 0.00 |
| *NUM/PNOUN* | 6 | 0.01 | 17 | 0.00 |
| *EXCL* | 3 | 0.00 | 10 | 0.00 |
| *VERB+PRN* | 7 | 0.01 | 9 | 0.00 |
| *ADJ/VERB* | 4 | 0.01 | 7 | 0.00 |
| *NOUN/EXCL* | 1 | 0.00 | 5 | 0.00 |
| *NOUN/PRN* | 2 | 0.00 | 3 | 0.00 |
| *ADJ/ADV/VERB* | 2 | 0.00 | 2 | 0.00 |
| *VERB/EXCL* | 1 | 0.00 | 2 | 0.00 |
| *ADJ/ADV/PNOUN* | 1 | 0.00 | 1 | 0.00 |
| *ADV/PRN* | 1 | 0.00 | 1 | 0.00 |
| *PNOUN/OTHER* | 1 | 0.00 | 1 | 0.00 |
| **TOTAL** | **67,802** | **100** | **1,379,521** | **100** |

*Note*: ADJ: Adjective, ADV: Adverb, ART: Article, CNJ: Conjunction, DM: Discourse Marker, EXCL: Exclamation, GER: Gerund, NOUN: Noun, NUM: Numeral, OTHER: Other, PNOUN: Proper Noun, PREP: Preposition, PRN: Pronoun, PTCL: Particle, VERB: Verb

**Table C2**

*Stress pattern distribution by major PoS – types*

| PoS | APU | PU | U | Other |
|---|---|---|---|---|
| *VERB* | 9,743 (47.35%) | 8,181 (39.76%) | 2,350 (11.42%) | 301 (1.46%) |
| *NOUN* | 5,815 (26.91%) | 9,946 (46.03%) | 4,887 (22.62%) | 958 (4.43%) |
| *ADJ* | 3,876 (37.08%) | 1,072 (10.26%) | 5,443 (52.08%) | 61 (0.58%) |
| *ADV* | 51 (14.29%) | 190 (53.22%) | 110 (30.81%) | 6 (1.68%) |

*Note*: "Other" includes unstressed or double stressed forms, and words with variant stress patterns.

**Table C3**

*Stress pattern distribution by major PoS – tokens*

| PoS | APU | PU | U | Other |
|---|---|---|---|---|
| *VERB* | 74,775 (35.31%) | 113,811 (53.75%) | 20,336 (9.60%) | 2,821 (1.33%) |
| *NOUN* | 79,680 (24.17%) | 162,281 (49.23%) | 80,395 (24.39%) | 7,288 (2.21%) |
| *ADJ* | 19,912 (30.85%) | 9,045 (14.01%) | 35,373 (54.81%) | 213 (0.33%) |
| *ADV* | 3,347 (11.10%) | 17,520 (58.10%) | 9,009 (29.87%) | 280 (0.93%) |

*Note*: "Other" includes unstressed or double stressed forms and words with variant stress patterns.

**Table C4**

*Cumulative stress pattern distribution in nouns – types*

| Grade Range | APU | PU | U | Other | Total Types |
|---|---|---|---|---|---|
| *1* | 709 (20.46%) | 1,866 (53.85%) | 787 (22.71%) | 103 (2.97%) | 3,306 |
| *1–2* | 1,110 (21.52%) | 2,746 (53.24%) | 1,148 (22.26%) | 154 (2.99%) | 6,775 |
| *1–3* | 2,091 (22.53%) | 4,729 (50.95%) | 2,140 (23.06%) | 322 (3.47%) | 17,445 |
| *1–4* | 3,619 (26.58%) | 6,450 (47.37%) | 3,079 (22.61%) | 469 (3.44%) | 32,686 |
| *1–5* | 5,095 (26.88%) | 8,780 (46.33%) | 4,286 (22.61%) | 792 (4.18%) | 57,581 |
| *1–6* | 5,815 (26.91%) | 9,946 (46.03%) | 4,887 (22.62%) | 958 4.43 | 79,680 |

*Note*: "Other" includes unstressed or double stressed forms and words with variant stress patterns.

**Table C5**

*Cumulative stress pattern distribution in nouns – tokens*

| Grade Range | APU | PU | U | Other | Total tokens |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| *1* | 3,306 (19.41%) | 9,043 (53.10%) | 4,390 (25.78%) | 292 (1.71%) | 17,031 |
| *1–2* | 6,775 (20.45%) | 17,531 (52.91%) | 8,177 (24.68%) | 648 (1.96%) | 33,131 |
| *1–3* | 17,445 (20.51%) | 44,268 (52.04%) | 21,543 (25.32%) | 1,812 (2.13%) | 85,068 |
| *1–4* | 32,686 (21.92%) | 75,680 (50.76%) | 37,394 (25.08%) | 3,344 (2.24%) | 149,104 |
| *1–5* | 57,581 (23.15%) | 124,895 (50.21%) | 60,626 (24.37%) | 5,637 (2.27%) | 248,739 |
| *1–6* | 79,680 (24.17%) | 162,281 (49.23%) | 80,395 (24.39%) | 7,288 (2.21%) | 329,644 |

*Note*: "Other" includes unstressed or double stressed forms and words with variant stress patterns.

**Table C6**

*Cumulative stress pattern distribution in verbs – types*

| Grade Range | APU | PU | U | Other | Types (*N*) |
|---|---|---|---|---|---|
| *1* | 1,257 (40.73%) | 1,446 (46.86%) | 320 (10.37%) | 63 (2.04%) | 3,086 |
| *1–2* | 1,892 (41.70%) | 2,079 (45.82%) | 486 (10.71%) | 80 (1.76%) | 4,537 |
| *1–3* | 3,945 (45.32%) | 3,671 (42.18%) | 920 (10.57%) | 168 (1.93%) | 8,704 |

| | | | | | |
|---|---|---|---|---|---|
| *1–4* | 5,903 (45.55%) | 5,370 (41.44%) | 1,489 (11.49%) | 196 (1.51%) | 12,958 |
| *1–5* | 8,544 (47.19%) | 7,265 (40.12%) | 2,028 (11.20%) | 269 (1.49%) | 18,106 |
| *1–6* | 9,743 (47.35%) | 8,181 (39.76%) | 2,350 (11.42%) | 301 (1.46%) | 20,575 |

*Note*: "Other" includes unstressed or double stressed forms and words with variant stress patterns.

**Table C7**

*Cumulative stress pattern distribution in verbs – tokens*

| **Grade Range** | **APU** | **PU** | **U** | **Other** | **Tokens (*N*)** |
|---|---|---|---|---|---|
| *1* | 3,751 (29.95%) | 7,449 (59.48%) | 983 (7.85%) | 341 (2.72%) | 12,524 |
| *1–2* | 7,329 (29.73%) | 14,438 (58.57%) | 2,348 (9.53%) | 534 (2.17%) | 24,649 |
| *1–3* | 21,175 (34.59%) | 33,638 (54.94%) | 5,261 (8.59%) | 1,148 (1.88%) | 61,222 |
| *1–4* | 36,929 (35.08%) | 57,300 (54.42%) | 9,506 (9.03%) | 1,550 (1.47%) | 105,285 |
| *1–5* | 59,305 (35.55%) | 89,636 (53.73%) | 15,567 (9.33%) | 2,324 (1.39%) | 166,832 |
| *1–6* | 74,775 (35.31%) | 113,811 (53.75%) | 20,336 (9.60%) | 2,821 (1.33%) | 211,743 |

*Note*: "Other" includes unstressed or double stressed forms and words with variant stress patterns.

**Table C8**

*Rank-ordered distribution of the top 10 word-level syllable templates by grade level, showing type counts and percentages of types*

| Grade | Template | Types (*N*) | Types (%) | Grade Type (*N*) | Rank |
|---|---|---|---|---|---|
| *1* | CV.ˈCV.CV. | 441 | 4.86 | 9,077 | 1 |
| *1* | ˈCV.CV. | 354 | 3.9 | 9,077 | 2 |
| *1* | ˈCV.CV.CV. | 264 | 2.91 | 9,077 | 3 |
| *1* | CV.CV.ˈCV.CV. | 193 | 2.13 | 9,077 | 4 |
| *1* | CV.ˈCV. | 188 | 2.07 | 9,077 | 5 |
| *1* | CV.ˈCV.CV.CV. | 166 | 1.83 | 9,077 | 6 |
| *1* | ˈCCV.CV. | 166 | 1.83 | 9,077 | 7 |
| *1* | ˈCV.CVC. | 141 | 1.55 | 9,077 | 8 |
| *1* | CCV.ˈCV.CV. | 130 | 1.43 | 9,077 | 9 |
| *1* | CV.CV.ˈCV. | 129 | 1.42 | 9,077 | 10 |
| *2* | CV.ˈCV.CV. | 481 | 4.16 | 11,575 | 1 |
| *2* | ˈCV.CV. | 416 | 3.59 | 11,575 | 2 |
| *2* | ˈCV.CV.CV. | 309 | 2.67 | 11,575 | 3 |
| *2* | CV.CV.ˈCV.CV. | 223 | 1.93 | 11,575 | 4 |
| *2* | CV.ˈCV. | 219 | 1.89 | 11,575 | 5 |
| *2* | CV.ˈCV.CV.CV. | 202 | 1.75 | 11,575 | 6 |
| *2* | ˈCCV.CV. | 201 | 1.74 | 11,575 | 7 |
| *2* | ˈCV.CVC. | 185 | 1.6 | 11,575 | 8 |
| *2* | CV.CV.ˈCV. | 164 | 1.42 | 11,575 | 9 |
| *2* | CCV.ˈCV.CV. | 150 | 1.3 | 11,575 | 10 |
| *3* | CV.ˈCV.CV. | 666 | 3.06 | 21,759 | 1 |
| *3* | ˈCV.CV. | 569 | 2.62 | 21,759 | 2 |
| *3* | ˈCV.CV.CV. | 518 | 2.38 | 21,759 | 3 |
| *3* | CV.ˈCV.CV.CV. | 370 | 1.7 | 21,759 | 4 |
| *3* | CV.CV.ˈCV.CV. | 351 | 1.61 | 21,759 | 5 |
| *3* | ˈCV.CVC. | 337 | 1.55 | 21,759 | 6 |
| *3* | ˈCCV.CV. | 291 | 1.34 | 21,759 | 7 |
| *3* | CV.ˈCV. | 252 | 1.16 | 21,759 | 8 |
| *3* | CV.CV.ˈCV. | 245 | 1.13 | 21,759 | 9 |
| *3* | CV.ˈCV.CVC. | 232 | 1.07 | 21,759 | 10 |

| | | | | | |
|---|---|---|---|---|---|
| *4* | CV.'CV.CV. | 846 | 2.54 | 33,322 | 1 |
| *4* | 'CV.CV.CV. | 679 | 2.04 | 33,322 | 2 |
| *4* | 'CV.CV. | 665 | 2 | 33,322 | 3 |
| *4* | CV.'CV.CV.CV. | 543 | 1.63 | 33,322 | 4 |
| *4* | CV.CV.'CV.CV. | 469 | 1.41 | 33,322 | 5 |
| *4* | 'CV.CVC. | 379 | 1.14 | 33,322 | 6 |
| *4* | 'CCV.CV. | 364 | 1.09 | 33,322 | 7 |
| *4* | CV.CV.'CV. | 355 | 1.07 | 33,322 | 8 |
| *4* | V.'CV.CV.CV. | 343 | 1.03 | 33,322 | 9 |
| *4* | CV.'CV. | 334 | 1 | 33,322 | 10 |
| *5* | CV.'CV.CV. | 1011 | 2.28 | 44,383 | 1 |
| *5* | 'CV.CV.CV. | 821 | 1.85 | 44,383 | 2 |
| *5* | 'CV.CV. | 773 | 1.74 | 44,383 | 3 |
| *5* | CV.'CV.CV.CV. | 704 | 1.59 | 44,383 | 4 |
| *5* | CV.CV.'CV.CV. | 581 | 1.31 | 44,383 | 5 |
| *5* | V.'CV.CV.CV. | 447 | 1.01 | 44,383 | 6 |
| *5* | 'CCV.CV. | 446 | 1 | 44,383 | 7 |
| *5* | 'CCV.CV.CV. | 442 | 1 | 44,383 | 8 |
| *5* | 'CV.CVC. | 435 | 0.98 | 44,383 | 9 |
| *5* | CV.CV.'CV. | 432 | 0.97 | 44,383 | 10 |
| *6* | CV.'CV.CV. | 1069 | 2.25 | 47,511 | 1 |
| *6* | 'CV.CV.CV. | 848 | 1.78 | 47,511 | 2 |
| *6* | 'CV.CV. | 784 | 1.65 | 47,511 | 3 |
| *6* | CV.'CV.CV.CV. | 745 | 1.57 | 47,511 | 4 |
| *6* | CV.CV.'CV.CV. | 637 | 1.34 | 47,511 | 5 |
| *6* | 'CV.CVC. | 491 | 1.03 | 47,511 | 6 |
| *6* | CV.CV.'CV. | 481 | 1.01 | 47,511 | 7 |
| *6* | V.'CV.CV.CV. | 474 | 1 | 47,511 | 8 |
| *6* | CV.'CV. | 457 | 0.96 | 47,511 | 9 |
| *6* | 'CCV.CV. | 454 | 0.96 | 47,511 | 10 |

**Table C9**

*Rank-ordered frequency distribution of the top 10 word-level syllable templates by grade level, showing token counts, corresponding types, percentages of tokens, and total grade token counts*

| Grade | Template | Tokens (*N*) | Types (*N*) | Tokens (%) | Grade Tokens (*N*) | Rank |
|---|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| *1* | CV. | 16,015 | 34 | 22.77 | 70,345 | 1 |
| *1* | 'CV.CV. | 3,354 | 354 | 4.77 | 70,345 | 2 |
| *1* | 'V.CV. | 2,847 | 62 | 4.05 | 70,345 | 3 |
| *1* | V. | 2,829 | 7 | 4.02 | 70,345 | 4 |
| *1* | CVC. | 2,503 | 14 | 3.56 | 70,345 | 5 |
| *1* | CV.'CV. | 1,935 | 188 | 2.75 | 70,345 | 6 |
| *1* | CV.'CV.CV. | 1,741 | 441 | 2.47 | 70,345 | 7 |
| *1* | CVN. | 1,611 | 13 | 2.29 | 70,345 | 8 |
| *1* | CCV. | 1,310 | 6 | 1.86 | 70,345 | 9 |
| *1* | V.'CV. | 1,205 | 22 | 1.71 | 70,345 | 10 |
| *2* | CV. | 21,704 | 34 | 22.64 | 95,886 | 1 |
| *2* | 'CV.CV. | 4,291 | 416 | 4.48 | 95,886 | 2 |
| *2* | 'V.CV. | 3,731 | 64 | 3.89 | 95,886 | 3 |
| *2* | CVC. | 3,286 | 12 | 3.43 | 95,886 | 4 |
| *2* | V. | 3,092 | 6 | 3.22 | 95,886 | 5 |
| *2* | CV.'CV. | 2,810 | 219 | 2.93 | 95,886 | 6 |
| *2* | CVN. | 2,092 | 13 | 2.18 | 95,886 | 7 |
| *2* | CV.'CV.CV. | 1,975 | 481 | 2.06 | 95,886 | 8 |
| *2* | CCV. | 1,725 | 8 | 1.8 | 95,886 | 9 |
| *2* | 'CV.CV.CV. | 1,715 | 309 | 1.79 | 95,886 | 10 |
| *3* | CV. | 49,998 | 38 | 21.73 | 230,041 | 1 |
| *3* | V. | 9,511 | 8 | 4.13 | 230,041 | 2 |
| *3* | CVC. | 9,370 | 21 | 4.07 | 230,041 | 3 |
| *3* | 'CV.CV. | 9,333 | 569 | 4.06 | 230,041 | 4 |

| | | | | | | |
|---|---|---|---|---|---|---|
| *3* | 'V.CV. | 7,696 | 89 | 3.35 | 230,041 | 5 |
| *3* | CVN. | 6,273 | 15 | 2.73 | 230,041 | 6 |
| *3* | CV.'CV. | 5,142 | 252 | 2.24 | 230,041 | 7 |
| *3* | V.'CV. | 4,305 | 34 | 1.87 | 230,041 | 8 |
| *3* | 'CV.CV.CV. | 4,115 | 518 | 1.79 | 230,041 | 9 |
| *3* | CV.'CV.CV. | 3,880 | 666 | 1.69 | 230,041 | 10 |
| *4* | CV. | 68,716 | 40 | 20.32 | 338,202 | 1 |
| *4* | CVC. | 13,598 | 22 | 4.02 | 338,202 | 2 |
| *4* | V. | 12,795 | 8 | 3.78 | 338,202 | 3 |
| *4* | 'CV.CV. | 12,551 | 665 | 3.71 | 338,202 | 4 |
| *4* | 'V.CV. | 11,535 | 101 | 3.41 | 338,202 | 5 |
| *4* | CVN. | 9,023 | 14 | 2.67 | 338,202 | 6 |
| *4* | CV.'CV. | 7,531 | 334 | 2.23 | 338,202 | 7 |
| *4* | V.'CV. | 6,018 | 45 | 1.78 | 338,202 | 8 |
| *4* | 'CV.CV.CV. | 5,711 | 679 | 1.69 | 338,202 | 9 |
| *4* | CV.'CV.CV. | 5,361 | 846 | 1.59 | 338,202 | 10 |
| *5* | CV. | 91,176 | 45 | 19.67 | 463,502 | 1 |
| *5* | CVC. | 18,952 | 30 | 4.09 | 463,502 | 2 |
| *5* | V. | 16,868 | 9 | 3.64 | 463,502 | 3 |
| *5* | 'CV.CV. | 16,148 | 773 | 3.48 | 463,502 | 4 |
| *5* | 'V.CV. | 15,974 | 120 | 3.45 | 463,502 | 5 |
| *5* | CVN. | 13,969 | 19 | 3.01 | 463,502 | 6 |
| *5* | CV.'CV. | 9,450 | 427 | 2.04 | 463,502 | 7 |
| *5* | V.'CV. | 8,374 | 64 | 1.81 | 463,502 | 8 |

| | | | | | | |
|---|---|---|---|---|---|---|
| *5* | 'CV.CV.CV. | 7,581 | 821 | 1.64 | 463,502 | 9 |
| *5* | 'CCV.CV. | 6,788 | 446 | 1.46 | 463,502 | 10 |
| *6* | CV. | 105,510 | 50 | 19.8 | 532,749 | 1 |
| *6* | CVC. | 20,825 | 36 | 3.91 | 532,749 | 2 |
| *6* | V. | 20,143 | 9 | 3.78 | 532,749 | 3 |
| *6* | 'V.CV. | 18,706 | 121 | 3.51 | 532,749 | 4 |
| *6* | 'CV.CV. | 17,995 | 784 | 3.38 | 532,749 | 5 |
| *6* | CVN. | 16,406 | 22 | 3.08 | 532,749 | 6 |
| *6* | CV.'CV. | 11,040 | 457 | 2.07 | 532,749 | 7 |
| *6* | V.'CV. | 9,633 | 65 | 1.81 | 532,749 | 8 |
| *6* | 'CV.CV.CV. | 8,084 | 848 | 1.52 | 532,749 | 9 |
| *6* | CCV. | 7,898 | 12 | 1.48 | 532,749 | 10 |