The CEOL Universe, version 1.0

An inventory of candidate English strings and the evidence for each

Ryan Kamp

Technical report. Draft of October 25, 2026. Every figure in it is printed by `pipeline/research/release/report/paper.py` from committed files.

Dataset	`ryanjosephkamp/ceol-universe` on the Hugging Face hub, version 1.0
Rows	10,020,150 candidates, 1 per string, in 7 configs
Built	October 25, 2026, by `pipeline/research/release/build_release.py` at commit `6891bff41ef0`
License	the data CC BY-SA 4.0; this report CC BY 4.0
Identifier	a DOI, minted at the release
Code and records	https://github.com/ryanjosephkamp/ceol-universe
Contact	ryanjosephkamp@gmail.com, for requests to withhold, correct, or restore a source or a row
Documentation	the dataset card (the dataset's `README.md`), the data dictionary, and `NOTICES.md`

Abstract

The CEOL Universe is a public inventory of 10,020,150 candidate English strings. Each row records which of 8 published word lists list the string, which lists cannot hold strings of its spelling class, whether 2 corpora contain it, and how it entered the pool. No row says that a string is a word. Of the 9,725,285 single-token rows, 1,089,968 (11.2%) are listed by at least 1 list and 8,033,152 (82.6%) rest on a corpus alone. Most listed rows are listed by 1 family: 43.7% of them by Wiktionary and by no other family. The inventory is a frame from which a reader derives a word list under a definition they state.

1. Introduction

Word lists disagree. A spell-checker's dictionary, a tournament's word list, a lexical database, and a corpus's vocabulary each draw the line around English somewhere else, and none is wrong by its own rule. A study that needs to know whether a string "exists" usually builds its own check against 1 reference, chosen for the occasion and not recorded with the result. The checks are hard to compare and hard to repeat.

The CEOL Universe takes the other approach. It keeps every candidate string its sources gave, under 1 executable spelling grammar, and records each source's answer beside the string, dated by the source's pinned version. A reader who wants a word list states a definition (for example, "listed by 2 or more independent families, single tokens only, no string a language model decided") and obtains a list, a count, and the code that reproduces both. The project's site turns that into an interface: a reader picks criteria, sees the count over all rows at once, and copies the code.

This report describes version 1.0, the first public version. Its contributions are:

- **The inventory:** 10,020,150 rows, each a candidate string with an identifier that depends on the string alone, so the same string has the same identifier in every version.
- **The evidence semantics:** a listing is 1 (listed), 0 (not listed), -1 (the list cannot hold strings of this class), or null (not measured). No figure in this report counts -1 or null as "not listed".
- **A comparison of the published lists over the same strings** (§7). Because every list is measured against every row, the report can state how far the lists contain one another, what each holds alone, and how listing depends on frequency, over the rows each list could hold.
- **The record of how each row came in** (§5 and §6), kept apart from the evidence for it.

What the universe is not. It is not a dictionary: it has no definitions, senses, or pronunciations. It does not decide whether a string is a word, and its project text never calls a candidate valid. It is not a sample of English text, and its counts are counts of strings, not of uses.

2. Related resources

This section sets the CEOL Universe beside the open English word lists and lexical databases a reader might compare it with, the documentation standards that bear on it, and what the literature says about contamination and canaries. Sizes and licenses are each resource's own statements, read on October 26, 2026, unless an entry gives another date. A string of the universe is a candidate; nothing here says whether any string is a word.

Comparable resources

Game-derived marks a list whose documentation, or its maintainer's documentation of the same source lists, says it takes its selection, in whole or in part, from a proprietary game word list: the American Official Scrabble Players Dictionary (OSPD), the British Official Scrabble Words (OSW), Collins Scrabble Words (CSW), which descends from both [1], or the Tournament Word List (TWL) [2]. The mark records what each list says about itself; it is not a finding about any entry.

Lexicons, Wiktionary, and its extractions

Resource	What it is	Size, as stated	License, as stated	What it records per entry	What it does not record
Princeton WordNet 3.0	A lexical database of English that groups nouns, verbs, adjectives, and adverbs into synonym sets (synsets) linked by conceptual-semantic and lexical relations. Princeton states that it is no longer developed and names Open English WordNet and the Global WordNet Association as community-led resources that build on it [3].	155,287 unique strings counted within each part of speech (147,278 distinct overall), 117,659 synsets, and 206,941 word-sense pairs [4].	The WordNet 3.0 license: use, copying, modification, and distribution for any purpose without fee, provided the copyright notice and statements appear on all copies; Princeton's name may not be used in publicity [5].	For each lemma and part of speech, its synsets, pointer (relation) types, and a count of tagged senses; for each synset, its words, lexicographer file, relations, and a gloss [6].	Parts of speech other than the 4 above [4]; the database format has no pronunciation or etymology field [6].
Open English WordNet, 2025 edition	A fork of Princeton WordNet developed under an open-source methodology, released in annual editions since 2019; from the 2025 release, proper nouns are in a separate resource, Open English Namenet [7].	2025 edition: 135,969 words and 107,519 synsets; 2025 Plus edition, which adds curated proper nouns: 161,875 words and 120,564 synsets [7].	CC BY 4.0 for the team's work, with the Princeton WordNet license applying to the underlying data; attribution to both is required [7].	Lexical entries (lemma and part of speech) and senses, grouped into synsets that carry a part of speech, an Interlingual Index link, relations, and examples [7]; definitions [8]; pronunciation for nearly 35,000 entries and Wikidata links for over 6,000, as of the 2021 release [8].	Proper nouns in the 2025 edition, which moved them to Open English Namenet [7].
English Wiktionary	A collaboratively edited dictionary whose pages hold one section per language; an entry is a single language section [9]. The text is an unstructured wiki maintained by volunteers [10].	English: 1,396,346 entries, 719,082 of them lemma entries, generated from the October 1, 2026 dump [9].	CC BY-SA 4.0 and the GNU Free Documentation License, under the Wikimedia Terms of Use in effect since June 7, 2023 [11].	Wikitext per language section; the next row lists what an extractor recovers from it [10].	A fixed schema: it is an unstructured wiki [10].

Resource	What it is	Size, as stated	License, as stated	What it records per entry	What it does not record
Wiktextract and kaikki.org	An extractor that expands Wiktionary's templates and Lua modules to produce structured JSON [10]; kaikki.org publishes its output [12].	English: 1,396,111 distinct words, extracted October 21, 2026 from the English Wiktionary dump dated October 1, 2026, as the page stated on October 26, 2026 [12].	The same licenses as Wiktionary, CC BY-SA and the GFDL, with no version stated [12].	Lemmas, inflected forms, translations, etymology, usage examples, pronunciations (with sound-file URLs), lexical and semantic relations, and morphological, syntactic, semantic, topical, and dialectal annotations [10].	A fixed release: the data is re-extracted regularly, usually at least once a week [12].
DBnary	Lexical data extracted from 27 Wiktionary language editions and published as linked open data in the OntoLex model, one version per Wikimedia dump [13, 14].	Not confirmed: the site was unreachable on October 26, 2026, and its home page states no counts [13].	Stated 2 ways: CC BY-SA 3.0 on its site [13]; CC BY 4.0 in its Zenodo record for version 20260601 [14].	Pages, entries, senses, translations, and lexical relations, with a set of disambiguated translations as an added layer [13].	Not stated in the pages read.
MorphyNet	A multilingual database of inflectional and derivational morphology drawn from Wiktionary editions [15].	English: 396,772 words with 649,594 inflected forms; 67,412 words with 225,131 derived forms [15].	CC BY-SA 3.0 [15].	Inflection rows (lemma, form, features, morpheme segmentation); derivation rows (source and target word, their parts of speech, the affix, and its type) [15].	Senses or pronunciation: its row formats have no such fields [15].
SPECIALIST Lexicon (National Library of Medicine)	A syntactic lexicon of biomedical and general English, designed for the UMLS SPECIALIST Natural Language Processing System and released annually since 1994 [16].	2027 release: 534,363 lexical items and 1,007,936 forms, 51.55% of the forms multiword [17].	NLM's terms, not a Creative Commons license: no charge; redistributions keep the conditions; a publication must name the release number and date and describe any modifications; no proprietary rights may be claimed [18].	Lexical records from which its tools retrieve syntactic category, inflectional and spelling variants, abbreviations, acronyms, and derivational variants; separate files of derivation pairs, synonyms, and antonyms [16].	Not stated in the pages read.
CMU Pronouncing Dictionary (CMUdict) 0.7b	A machine-readable pronunciation dictionary for North American English, maintained by Carnegie Mellon University's Speech Group [19, 20].	"Over 134,000 words and their pronunciations" [19].	A license file, Copyright 1993–2015 Carnegie Mellon University, that permits redistribution and use with or without modification if the copyright notice and disclaimer are kept [20]; the README calls research and commercial use completely unrestricted and asks users to acknowledge its origin [20].	A pronunciation in the ARPAbet set of 39 phonemes, with a lexical stress marker on each vowel [19].	Anything but the word and its pronunciation: its entries map words to pronunciations [19].

Word lists

Resource	What it is	Size, as stated	License, as stated	What it records per entry	What it does not record
SCOWL (SCOWLv1), release 2020.12.07. Partly game-derived at sizes 80 and above (see the last column).	Spell Checking Oriented Word Lists: word lists split by size and spelling category, intended for use in spell checkers [21]. As of February 17, 2026, the original format is no longer maintained [22].	The SCOWLv1 README (commit of July 22, 2024) counts 658,213 entries through size 95 in the American and English categories, names included [21].	Copyright 2000–2018 Kevin Atkinson: permission to use, copy, modify, distribute, and sell, provided the notice appears in all copies; the notices of its components also apply, among them UKACD's, which requires that its notice be displayed and its text included verbatim on redistribution [23].	Membership by spelling category (American, British, Canadian, Australian, and variant levels), sub-category (abbreviations, contractions, proper names, upper case, words), and size (10 to 95) [21].	Hyphenated and open (spaced) compounds, which the original version excluded [24]. Note: size 80 includes all the lists of the ENABLE supplement package except one [23], and SCOWL's successor documents that package's lists of additional and stale OSPD words [25].
English Speller Database (ESDB, formerly SCOWLv2). Partly game-derived at sizes 80 and 85 (see the last column).	SCOWL's successor: a SQLite database and text file on how common a word is, dialect spelling differences, variants, and basic part of speech and inflection [24]. It is a work in progress, and the database itself will not have an official release until it stabilizes [26].	No entry count stated; the minimum size is 35, and size 95 is not included [24].	Copyright 2000–2026 Kevin Atkinson, the same notice-retention grant, which the README calls "MIT-like"; further notices apply to Australian output, to generated lists larger than size 80 (UKACD), and possibly WordNet [24, 27].	Size, region, spelling and variant level, a lemma with its part of speech, inflected forms, definition notes, usage notes (among them offensive and vulgar), and sometimes a tag naming the source list [24]. Unlike SCOWLv1, it includes hyphenated and open compounds [24].	Note: its sources document lists additional OSPD words, added at size 80, and stale OSPD words, at size 85, both from the ENABLE2K supplement files [25].
LibreOffice hunspell dictionaries en_US, en_CA, en_AU, en_GB	Spell-checker dictionaries. The en_US, en_CA, and en_AU README (version 2020.12.07) says these dictionaries are made by a SCOWL script and that the normal ones correspond to SCOWL size 60; en_GB began from a subset of Kevin Atkinson's list and has been extensively updated by others [28].	No entry count stated in the READMEs [28].	en_US, en_CA, en_AU: they come directly from SCOWL and are under SCOWL's copyright, with an affix file under Geoff Kuenning's BSD license; en_GB: "covered by his original LGPL licence", with the affix file also under the LGPL [28].	Words and affix rules that generate inflected forms, 1 spelling variety per dictionary; en_GB adds hyphenated compounds, place names, possessives, plurals, and a morphological extension based on Atkinson's part-of-speech and inflection data [28].	Most alternative spellings: the size-60 dictionaries generally include 1 variant per word [28].

Resource	What it is	Size, as stated	License, as stated	What it records per entry	What it does not record
ENABLE (ENABLE2K). A word-game list; it states it is not the tournament list.	The Enhanced North American Benchmark Lexicon, a word list for word games researched and compiled by Alan Beale. Its README says the master list is not the proprietary TWL98, and that it reflects the OSPD changes in effect from January 1, 1998 [29].	173,528 words in the Y2K update; the first edition had 100,701 [29].	The master list is "formally released into the Public Domain"; credit to the originators is requested [29].	Membership in a plain ASCII word list [29].	Anything beyond the string [29].
12dicts 6.0.2 (June 2016). Partly game-derived (some inflections in 2of12inf).	Alan Beale's word lists of common words; the classic lists record agreement among 12 source dictionaries of American English, 8 for learners of English and 4 desk dictionaries [30].	Classic lists from about 22,000 to about 82,000 words [30].	Released to the public domain, with acknowledgment requested; 2of12inf and the 2+2+3 lists depend on AGID, which keeps them out of the public domain [30].	Membership in lists defined by how many source dictionaries hold a word; 2of12inf adds inflections, marking the plurals of uncountable nouns [30].	Note: the README says some 2of12inf inflections came from Scrabble-related sources, and that no later list with inflections includes those plurals [30].
Moby Words II (Grady Ward). Game-derived: the 2 crossword files.	Word lists by Grady Ward, released on Project Gutenberg on May 1, 2002 [31].	Separate lists, among them 354,984 single words, 256,772 compound words, 74,550 common dictionary words, 21,986 names, 10,196 places, and 113,809 crossword-game words [31].	"Public Domain material by grant from the author, January, 2001" [31].	Membership in each list [31].	Accents, which were stripped. The crossword file is described as compatible with the first edition of the OSPD, and a delta file makes it compatible with the second [31].
Moby Part-of-Speech II (Grady Ward)	A part-of-speech list, the second edition, with about 15,000 new entries [32].	No count stated in its documentation [32].	The same 2001 public-domain grant [32].	A word or phrase with 1 or more of 15 part-of-speech codes, the principal use first [32].	Many -ed, -ing, -ly, and -ic forms, which are not listed explicitly [32].
YAWL (Yet Another Word List) 0.3.2.03. Game-derived.	A word-game list by M. Leo Cooper, a superset of ENABLE, OSW, and lists by Alan Beale; rehosted on GitHub by Aaron Bull Schaefer [33].	264,097 words [33].	The list "is in the Public Domain" [33].	Membership [33].	Anything beyond the string [33].
Letterpress word list. Game-derived.	The word list behind the Letterpress game, loosely based on other word lists; its rule is that a word allowed in another popular word game such as Scrabble should be allowed, and it names YAWL, ENABLE2K, and CSW12 as useful lists [34].	Not stated [34].	CC0 1.0 Universal [34].	Membership in 1 sorted file per language; US, UK, and Australian spellings combined [34].	Proper nouns and acronyms, excluded by rule with a few kept for historical reasons; which spelling variety an entry belongs to [34].

Resource	What it is	Size, as stated	License, as stated	What it records per entry	What it does not record
Wordnik wordlist. A word-game list; the source of its selection is not stated.	A list of English words commonly used in word games, for game developers [35]. Wordnik's paid game dataset, which the README offers to those who need more data, states that it takes its playable words from ENABLE and the Tournament Word List [2].	Not stated [35].	MIT License, Copyright 2020 Wordnik [35].	Membership [35].	Definitions, parts of speech, frequency, and offensive-term labels, which the README offers in a licensed dataset [35].
dwyl/english-words	A text file of English words extracted from a word list once published by Infochimps [36].	"Over 466k" words in the README; 479k in the repository's description [36].	The Unlicense, beside a README that says the copyright still belongs to the original publisher [36].	1 string per line; a second file keeps only letters-only strings [36].	Anything beyond the string [36].
Webster's Second word lists (web2, web2a)	Word lists in FreeBSD's `share/dict` directory, introduced as Webster's Second International; web2a holds hyphenated terms and noun and adverbial phrases. FreeBSD maintains web2 as a list of current American English spellings and has removed a few entries [37].	web2: 234,936 words; no count for web2a [37].	No license stated; the README says the 1934 copyright "has lapsed, according to the supplier" [37]. The Catalog of Copyright Entries records renewals of the dictionary's 1934 copyrights on December 18, 1961 [38].	Membership [37].	Anything beyond the string [37].

Frequency resources and norms

Resource	What it is	Size, as stated	License, as stated	What it records per entry	What it does not record
wordfreq 3.1.1 (Robyn Speer)	A Python library of word frequencies in over 40 languages, combining several sources; its data is a snapshot of usage through about 2021 and is unlikely to be updated again [39, 40]. 3.1.1, uploaded November 21, 2023, is the newest release on PyPI [41].	No English word count stated; its table lists 7 sources for English [39].	Code under the Apache license; data files that may be redistributed under CC BY-SA 4.0 [39]. The current README adds that converting the data to CSV does not follow the license, because CSV has no place for attribution [39].	A frequency per word, also on the Zipf scale (the base-10 logarithm of occurrences per billion words) [39].	Words below its cutoff: the lowest Zipf value in the large lists is 1.0, and a word not in the list returns a default value of 0, not a missing value [39].

Resource	What it is	Size, as stated	License, as stated	What it records per entry	What it does not record
Google Books Ngram, English, version 20200217	N-gram counts from the books Google has scanned; the downloadable sets were generated in February 2020 (version 3), July 2012 (version 2), and July 2009 (version 1) [42].	No English 1-gram type count stated [42]. The first corpus held 5,195,769 books, with 361 billion English words [43]; the 2012 edition held 8,116,746 books, the English part close to half a trillion words [44].	"This compilation is licensed under a Creative Commons Attribution 3.0 Unported License" [42]; the Viewer's page says its graphs and data may be freely used for any purpose, with acknowledgment appreciated [45].	Per n-gram and year, a match count and a volume count (the version 2 format) [42].	N-grams that occur 40 times or fewer across the corpus, which versions 1 and 2 omit [42]; how often a book was read, since the corpus holds one copy of each book [46].
SUBTLEX-US	Word frequencies from American film and television subtitles [47].	74,286 words in the corpus, 60,384 of them with a frequency above 1 [48], from 51.0 million words in 8,388 films and episodes [47].	The 2009 paper: freely available for research purposes [47]. The author's copy on OSF carries a license file dated May 7, 2024 stating CC BY-NC-SA 4.0 [49]. The Ghent University download page states no license [48].	A frequency count, the number of films a word occurs in, a count of lowercase occurrences, and per-million values [48].	Not stated in the pages read.
SUBTLEX-UK	British English word frequencies from BBC subtitles broadcast from January 2010 to December 2012 [50].	332,987 word types in 201.7 million tokens; the main file holds 160,022 types (159,235 single words and 787 hyphenated) [50].	The author's OSF copy carries a license file dated May 7, 2024 stating CC BY-NC-SA 4.0 [51].	Frequency, Zipf value, contextual diversity, the dominant part of speech and its lemma, and all parts of speech observed [50].	Types seen in a single broadcast, and types with digits or other non-letter symbols, which the main file leaves out [50].
English Lexicon Project	Behavioral and descriptive data for lexical decision and naming studies [52].	40,481 words and 40,481 nonwords [52].	No terms on the home page (archived October 14, 2026); an About page archived in 2019 says word lists from the site are for non-commercial research only and may not be used to develop speech technology [53]. The live site's certificate had expired when read on October 26, 2026.	Lexical decision and naming responses (816 and 444 participants) and descriptive lexical statistics [52].	Not stated in the pages read.
Web 1T 5-gram Version 1 (LDC2006T13). Not open.	English word n-grams and their counts from about 1 trillion tokens of text from publicly accessible web pages, contributed by Google [54].	13,588,391 unigrams; 1,024,908,267,229 tokens [54].	The Web 1T 5-gram Version 1 Agreement of the Linguistic Data Consortium; fees apply [54].	N-grams of 1 to 5 words with observed frequency counts [54].	Not stated in the pages read.

Documentation standards

Each standard below asks a dataset to state certain things. The table gives the version read and the points that bear most on a candidate inventory.

Standard	What it asks a dataset to state
Datasheets for Datasets (version 8, 2021)	Questions in 7 sections: motivation, composition, collection process, preprocessing and labeling, uses, distribution, and maintenance [55]. Composition asks whether the data contains anything that might be offensive or insulting; distribution asks whether the dataset has a DOI; maintenance asks for an erratum, whether older versions stay available, and how others can contribute [55]. The questions are not prescriptive, those that do not apply may be skipped, and the process is not meant to be automated [55].
Data Statements (schema versions 1, 2, and 3)	Version 1 has 9 elements, among them curation rationale, language variety (a BCP-47 tag with a prose description), speaker and annotator demographics, and a provenance appendix carrying the statements of source datasets; a short form of about 60 to 100 words should appear in any publication that uses the dataset to train, tune, or test a system [56]. Version 2 has 15 elements, adding a header, an executive summary of 60 to 100 words, limitations, metadata with errata, and disclosures; it says missing information should be marked as missing, not used as a reason to skip the statement [57]. Version 3 has 17 elements, adding distribution and maintenance [58].
Data Cards (2022)	31 themes, among them the dataset's publishers and funding, updates and versions, differences across versions, upstream sources, inclusion and exclusion criteria, typical and outlier examples, and known patterns and biases, asked at 3 levels of detail [59].
Hugging Face dataset card template	Sections, among them dataset details, uses and out-of-scope uses, structure, creation (curation rationale, source data, and optional annotations), bias, risks, and limitations, citation, and contact [60]. It has no section for versions, a change log, or maintenance, and its personal and sensitive information section sits inside the optional annotations section [60].
Croissant 1.0 and 1.1, with the Responsible AI extension 1.0	A machine-readable metadata format for datasets [61], built on schema.org properties [62]. Version 1.0 was published on March 1, 2024 [63]. Version 1.1 (January 29, 2026) requires 9 properties, among them license, creator, and publication date; recommends a version number following MAJOR.MINOR.PATCH; strongly recommends a sha256 checksum for each file of a versioned dataset; and recommends the W3C PROV Ontology for provenance [62]. The Responsible AI extension (March 6, 2024) defines 20 properties, among them collection time frame, limitations, biases, use cases, personal and sensitive information, and a maintenance plan [64]. The NeurIPS 2026 Evaluations and Datasets track asks dataset authors for both the core and the Responsible AI fields [65].
FAIR Guiding Principles (2016)	15 principles, among them a persistent identifier, metadata that stay accessible after the data are gone, a clear data usage license, detailed provenance, and domain community standards [66].

Lexicographic standards. The TEI Lex-0 guidelines (version 0.9.5, February 8, 2026) require a header with a file description and a profile description, a license recorded inside the metadata, and a description of the sources the resource was derived from, and they call a recorded revision history good practice [67]. The ELEXIS Lexicographic Data Seal of Compliance (2021) assesses 6 principles: trustworthiness, interoperability, stewardship, and citability for the base seal, and reciprocity and openness for the higher one [68]. Its trustworthiness criteria ask for the language varieties covered and the rights holder of the dataset and of its source; its citability criteria ask how to cite the whole and a single entry; its reciprocity criteria ask whether users can submit corrections [68]. The European Language

Grid's minimal metadata ask, for a lexical resource, its basic unit of description and the kinds of information attached to it, and, for all data, language coverage and size and format per distribution [69]. The LRE Map asked authors at LREC conferences to describe each resource they built or used by its type, language, modality, use, availability, maturity, size, license, and documentation [70].

Documenting large corpora and citing evolving data. Documentation of the C4 web corpus examined where its text came from, what it contains, and what its filters removed, and found evaluation examples from other benchmarks inside it [71]. What's In My Big Data runs 16 analyses over 10 English corpora, benchmark contamination among them [72]. The Research Data Alliance's recommendations for citing evolving data ask for versioned and timestamped data, a timestamp and a checksum for each query result, and persistent identifiers for the queries [73]. The Tromsø recommendations for linguistic data ask a citation for author, date, title, publisher, and locator, and for the version where a dataset has versions [74].

None of the standards above asks a dataset to state its exposure to contamination or to carry a canary: a text search for both terms in each standard read (Datasheets, the 3 Data Statements schemas, Data Cards, the Hugging Face template, Croissant and its Responsible AI extension, FAIR, TEI Lex-0, the Seal of Compliance, the European Language Grid and META-SHARE schemas [75], the LRE Map, and the 2 citation recommendations) found neither.

Contamination and canaries

What contamination is. Sainz and colleagues separate guideline contamination (a model has seen a benchmark's annotation guidelines), raw-text contamination (it has seen the original text before annotation), and annotation contamination (it has seen the labels) [76]. Xu and colleagues grade contamination in 4 levels, from semantic (content on the same topic or from the same source) to label level [77].

What detection can show. Deng and colleagues organize detection methods into retrieval, temporal cutoff, masking, perturbation, canonical order, behavior manipulation, and membership inference [78]. Min-K% Prob, a method that needs no reference model, reaches an average AUC of 0.72 on the WikiMIA benchmark [79]. Guided instruction flags an instance when a model, given a dataset's name and the start of an instance, completes it exactly or nearly so [80]. The canonical-order test's false-positive guarantee rests on the dataset's order being exchangeable [81]. GPT-3's authors flagged test examples with a 13-gram overlap with the training data, and a bug left some overlaps in place [82]. Memorization grows log-linearly with model capacity, with how often an example is duplicated, and with the length of the prompt [83]. Xu and colleagues conclude that large-scale pretraining makes it nearly impossible to eliminate the risk fully [77].

Canaries. A canary in this sense is a marker placed in published material. BIG-bench puts a GUID canary in every task file so that researchers can filter its data out of web-scraped corpora, and so that a model can be tested afterward for training on it; it asks publications that quote its tasks to carry the string [84]. GPQA and BrowseComp each embed a canary that contains the BIG-bench string, to make their data easier to filter out of training corpora [85, 86], and a catalog of questionable practices in machine learning suggests that new benchmarks reuse the exact string [87]. Some developers report filtering on canaries. Anthropic's system card for Claude Opus 4.5 states that it dropped training documents containing the BIG-bench or Alignment Research Center canaries, reports that rephrased evaluation material persisted despite its filters, and suggests that writers of public evaluations attach canaries [88]. The PaLM report states that the BIG-bench canary does not appear in PaLM's training data [89]. The GPT-4 report says portions of BIG-bench were mixed into training and that it excluded BIG-bench from its results; it does not mention a canary [90].

What a canary cannot do. It works only where trainers know of it and respect it, and it does nothing against a developer who means to use the data [91]. It can be removed: third-party copies of GPQA and Humanity's Last Exam on Hugging Face were found with the canary sliced out [92], and removal is a stated limit of the method [93]. A memorization canary is a different object: the Secret Sharer study inserts random sequences into training data a chosen number of times and measures their exposure [94].

Other measures. Jacovi and colleagues propose encrypting public test data with a public key under a license that forbids derivatives, refusing to send test data to closed APIs that offer no training exclusion, and avoiding data that appears online with its solution [95]. Deng and colleagues add benchmarks built from recent text, dynamic refreshing, and protection of ground-truth labels [78]. Sainz and colleagues call for a registry of contamination cases [76]; the 2024 CONDA shared task's database held 566 entries over 91 contaminated sources from 23 contributors [96].

Dating a release. Several benchmarks make contamination questions answerable by date. LiveCodeBench tags each problem with its release date and, for a newer model, scores only the problems released after the model's cutoff [97]. LiveBench adds questions monthly and holds each month's new questions back for a month [98]. NewTerm collects the terms that the Cambridge, Collins, and Oxford online dictionaries add each year and plans annual updates [99]. WikiMIA uses Wikipedia timestamps and model release dates to separate text a model could have seen from text it could not [79]. Deng and colleagues caution that a time cutoff works only if data before and after it differ, since later text can repeat earlier text [78]. Resources date themselves too: wordfreq describes its data as a snapshot through about 2021 and gives text generated by language models as a reason it will not be updated [40].

How the universe applies this. The release's date and file hashes are recorded so that a later contamination question can be dated [100]. The dataset card carries a canary string, also recorded in the release record and in each Parquet file's metadata, so that a later test can check whether a model has memorized the dataset's documentation [100]. Each published file's sha256 is on the card [100], as Croissant recommends for versioned datasets [62].

What the universe adds, and what it lacks

What it records. For each candidate, the universe records its status in each of 8 published lists: WordNet 3.0, the English sections of a dated Wiktionary dump, the 4 LibreOffice hunspell dictionaries, ENABLE, and SCOWL [100]. For the 7 lists recorded as yes or no, each value is 1 (listed), 0 (not listed), -1 (the list cannot hold strings of that class, as a list without hyphens cannot hold a hyphenated string), or null (not measured); for SCOWL, the value is the size of the smallest SCOWL list that holds the string, 0 where none does, -1 where the class does not apply, and null where not measured [100]. It also records whether a string is a type of the 2020 Google Books English 1-grams and whether wordfreq 3.1.1 scores it [100]; the intake channel by which the string entered the pool, kept apart from containment [100]; and whether the English OpenList decided a row of its valid list with a language model [100]. Its candidates are the output of an executable grammar that admits hyphenated compounds, accented loanwords, contractions, apostrophe forms, ligature and diaeresis spellings, and multiword terms [100]. Each row's identifier is derived from the string alone, so the same string carries the same identifier in every version [100].

Against the resources above.

- *Per-list listing.* The word lists above record their own membership. The nearest to a per-source record are ESDB, whose tags sometimes name the source list a word came from [24], and Wordnik's paid game dataset, which labels words by source [2].
- *Not measured is kept apart from not listed.* wordfreq returns a default value of 0 for a word it does not hold [39]; in the universe, null means not measured by that source, never false [100].
- *Dated, fixed listings.* Several comparable resources change between readings: kaikki.org re-extracts its data usually at least weekly [12], DBnary makes one version per Wikimedia dump [14], and Open English WordNet releases annual editions [7]. The universe's listings are fixed at pinned versions, and each published file's hash is on its card [100].
- *Strings with spaces and hyphens.* SCOWLv1 excluded hyphenated and open compounds [24]; ESDB [24], Moby's compound list [31], and web2a [37] hold some. The universe states the grammar it admits strings under and marks each class [100].
- *Existence checks in benchmarks.* Benchmarks of invented terms check existence when they are built and drop what the check finds: PhantomBench keeps a concept only if it has zero

exact matches in the Dolma v1.7 corpus [101]; HypoTermQA removed every generated term with a web search result, keeping 790 of 1,000 [102]; and a vocabulary-test benchmark excluded 2 Spanish nonwords after finding them in dictionaries [103]. The universe records, per string, which published lists hold it and which were not measured, as a stored record rather than a filter [100].

What it lacks. The universe records no definitions or senses, which WordNet [6], Open English WordNet [8], Wiktionary and its extractions [10], and DBnary [13] record. It records no part of speech, though a candidate listed by WordNet could take one from WordNet [100]; WordNet [6], Open English WordNet [7], ESDB [24], Moby Part-of-Speech II [32], the SPECIALIST Lexicon [16], MorphyNet [15], and SUBTLEX-UK [50] each record one. It records no pronunciation, which CMUdict [19], Wiktextract [10], and part of Open English WordNet [8] record, and no etymology, which Wiktextract records [10]. It records no lemma or inflection relations [100], which ESDB [24], MorphyNet [15], and the SPECIALIST Lexicon [16] record. It records a count from the Google Books n-grams for most single-token rows, a corpus of scanned books, but a frequency drawn from many kinds of text, wordfreq's, for few rows [100]; wordfreq [39], the Google Books n-grams [42], and SUBTLEX [48, 50] supply frequencies. It records no behavioral measures, which the English Lexicon Project provides [52]; no spelling-variant levels, which SCOWL and ESDB record [21, 24]; and no variety labels, only 2 proxies from the hunspell dictionaries [100]. It records no validity: no row says that a string is a word [100].

3. Design

The candidate

A row of the universe is a **candidate**: a string that passes the spelling grammar, with the record of where it came from. The grammar (`pipeline/ceol/grammar.py`) normalizes a source entry to lower case, canonicalizes the typographic apostrophe and the no-break space, and admits it when every character is in a charset of 50 characters: the 26 letters a to z, 19 accented letters, 2 ligatures, the hyphen, the apostrophe, and the space. A space separates tokens, so a multiword term is 1 candidate.

The grammar marks each candidate with its spelling classes. A candidate with no class is **plain**; the classes are `accented`, `apostrophe-form`, `contraction`, `diaeresis`, `hyphenated`, `ligature`, `s-final` (a form ending in an apostrophe and s), and `spaced`. A candidate can carry several. The site's Definitions section runs the grammar on any typed string.

The evidence semantics

For each published list, a single-token row holds 1 where the list lists the string, 0 where it does not, and -1 where the list is structurally blind to the row's class (a list built without hyphens cannot hold a hyphenated string). A multiword row is measured by the 2 lists that hold multiword terms, WordNet and Wiktionary, and holds null in every other list's column. Null means "not measured by that source", never "false". Every share in this report is computed over the rows the list in question could hold, and says so.

The corpora contribute containment and features and never admit a string: Google Books' English 1-grams and wordfreq's English list each record whether they hold a single-token row, and the feature layer records how often.

Families and tiers

The 8 published lists come from 5 source families: WordNet; Wiktionary's English pages; hunspell, whose 4 English dictionaries (American, British, Canadian, and Australian) derive from 1 upstream and count as 1 family; ENABLE; and SCOWL. A family lists a row when any of its lists does, and cannot hold it when none of its lists can. Independence is counted by family, so a row that SCOWL and 2 hunspell dictionaries list counts as listed by 2 families.

Counting families this way follows the analysis plan's definition of the public panel. It is generous to SCOWL and hunspell: the American, Canadian, and Australian hunspell dictionaries come from SCOWL, as their READMEs state. Counted as 1 family, SCOWL and hunspell leave 372,809 rows (3.8%) listed by 2 or more families, 45,055 fewer than the count below. Every figure that counts families states which count it uses; the rest of this report uses the plan's.

2 of the published lists also draw in part on word-game material, by their own documentation. ENABLE's README says it reflects the changes to the Official Scrabble Players Dictionary in effect from January 1, 1998 [29], and SCOWL's lists at size 80 and above include the lists of the ENABLE supplement package [23], which SCOWL's successor documents as holding additional and stale entries of that dictionary [25]. Where those lists agree with a withheld game list, the agreement is not independent evidence.

Each single-token row falls in 1 of 4 **evidence tiers**: listed by 2 or more families; by 1; by none, with a corpus holding it ("corpus only"); or by none and no corpus ("no evidence"). The tiers are shown, not averaged away, because most rows are in 1 tier.

Evidence tier	Rows	Share of single-token rows
2+ families	417,864	4.3%
1 family	672,104	6.9%

Evidence tier	Rows	Share of single-token rows
corpus only	8,033,152	82.6%
no evidence	602,165	6.2%

The identifier and the canary

`id` is the first 16 hexadecimal digits of the sha256 of the string's UTF-8 bytes. It depends on the string alone, so a reader can join versions by `id` or by `word`, and the `sample` config, the rows whose `id` begins with `00`, holds the same strings in every version that holds them. The card and each data file carry a canary string, recorded in the release record, so that a later test can ask whether a model has memorized the dataset's documentation.

4. Sources, licenses, and the public projection

The build measured 13 panel sets. Each source's terms were read at its primary source, and the author ruled on the readings. 5 lists are withheld from the public version, because their membership may not be published: NASPA Word List 2023 and Collins Scrabble Words 2021, which are proprietary; the legacy SOWPODS list, for which no license was found; and Webster's Second International and its supplement of hyphenated terms, whose 1961 copyright renewal holds. Their columns and labels are absent. Their rows still ship wherever published evidence supports them.

The withholding has measured effects, which the card states and this report repeats:

- **Restricted rows.** 1,985 of the pin's 10,296,606 candidates are attested only by the 2 proprietary lists and are in no version.
- **Held-back rows.** Of the 10,294,621 shippable rows, 274,471 are held back in the private record, because once the withheld lists and channels are taken away nothing published supports them.

Why the row is held back	Rows
every listing a withheld list, and no public evidence left	27,661
a withheld channel label its only evidence	244,252
a census or audit row whose every source is a withheld list	2,558

- **Recomputed fields.** `panel_sets`, `panel_count`, and `provenance_class` count the 8 published lists alone, so a row with `panel_count` 0 may still be listed by a withheld list.
- **Withheld channel labels.** 210,109 rows carry `eol:channel-withheld`: every intake channel recorded for them names a withheld list or a source whose terms do not allow publishing the label.

- **Withheld feature columns.** 12 of the feature layer's columns are withheld. 10 were computed against a reference that holds Webster's Second, so a value could reveal whether that list holds a string's neighbor, stem, or segments. The other 2 come from a character model fitted on strings that at least 2/3 of their applicable panel sets list, the withheld lists among them; a test showed that their values, less those of the same model fitted on the published lists alone, rank strings by withheld listing well above chance (`docs/licenses/FEATURE_COLUMNS.md`).

The universe is licensed CC BY-SA 4.0, the share-alike floor of its Wiktionary-derived fields. `NOTICES.md` carries every notice and credit the sources require.

5. Construction

Each step has a committed record, and the counts are those records'. The site's "How it was built" page shows the same steps with their counts.

1. **The English OpenList's universe.** The English OpenList (EOL) is a word list with its own grammar and governance. Its valid and invalid lists at revision `c1139698` hold 9,653,999 strings, every string it had taken in, whichever way it ruled.
2. **The census.** The pinned sources re-read under the grammar give 213,867 candidates in the exception classes the EOL's letters-only rule could not hold: hyphens, accents, a diaeresis, apostrophes, and ligatures.
3. **The merge.** The 2 sets merged give 9,867,866 rows, 1 per string. Filtering the merge by its EOL flag reproduces the EOL's pinned lists byte for byte.
4. **The audit.** Every source's coverage by the merged universe was reconciled to a remainder of 0: 133,875 strings a source listed and the merge lacked were added. The result is the single-token pin, 10,001,741 candidates.
5. **The spaced admission.** 294,865 multiword candidates from WordNet and Wiktionary were admitted under the grammar's space rule. The pin, 10,296,606 rows, reproduces the single-token pin byte for byte as its single-token projection.
6. **The layers.** No row is added. Each row is measured against the panel lists (the evidence matrix), given its corpus features (the feature layer), and given its record of how it came in (the intake-provenance layer).
7. **The model-and-engine column,** derived from the EOL's valid dictionary at the pin (345,103 entries).
8. **The release.** The restricted rows are dropped and the public projection applied (§4), leaving 10,020,150 rows. The build asserted, on each of its 7 written files, that none of the 1,985 restricted rows' strings is present, that every `id` is distinct, that each filter config is its

predicate's filter, that no withheld column, label, or held-back row is present, and that `provenance_class` and `model_or_engine` re-derive from the file's own fields.

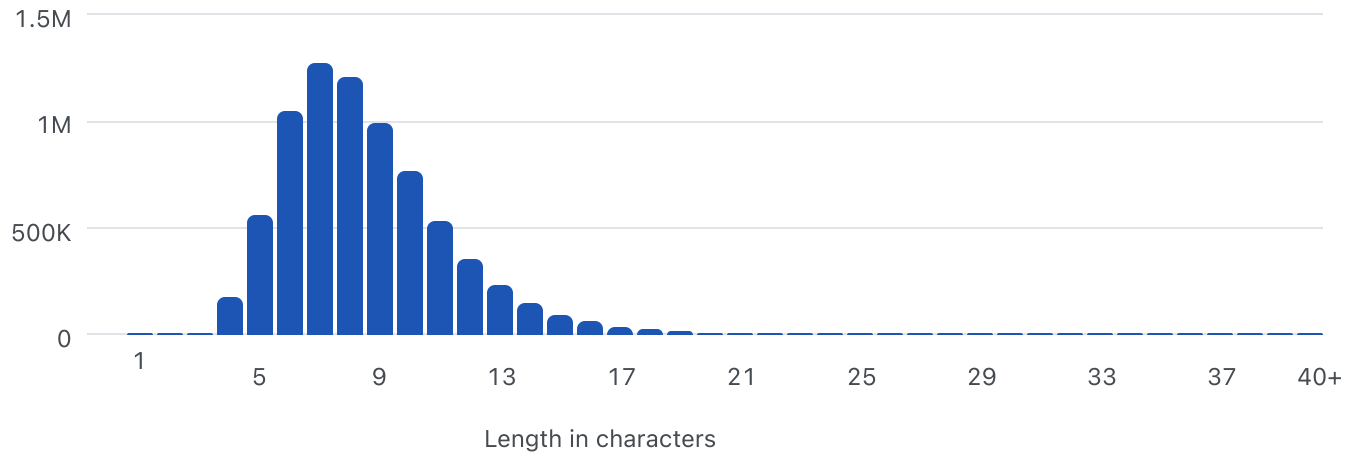
The release record, `pipeline/research/release/release_v1.0.json`, holds every file's rows, columns, bytes, and sha256, and the counts of every assertion.

6. Contents

Form

Length in characters, by evidence tier

The rows of each group by their length; the last column holds every longer row.



Shown: corpus only

Length. The median single-token row has 8 characters in the corpus-only tier, 9 in the tier listed by 2 or more families, 10 in the tier listed by 1, and 11 in the tier with no evidence. Multiword rows have a median of 14 characters; 229,297 of the 294,865 multiword rows (77.8%) have 2 tokens.

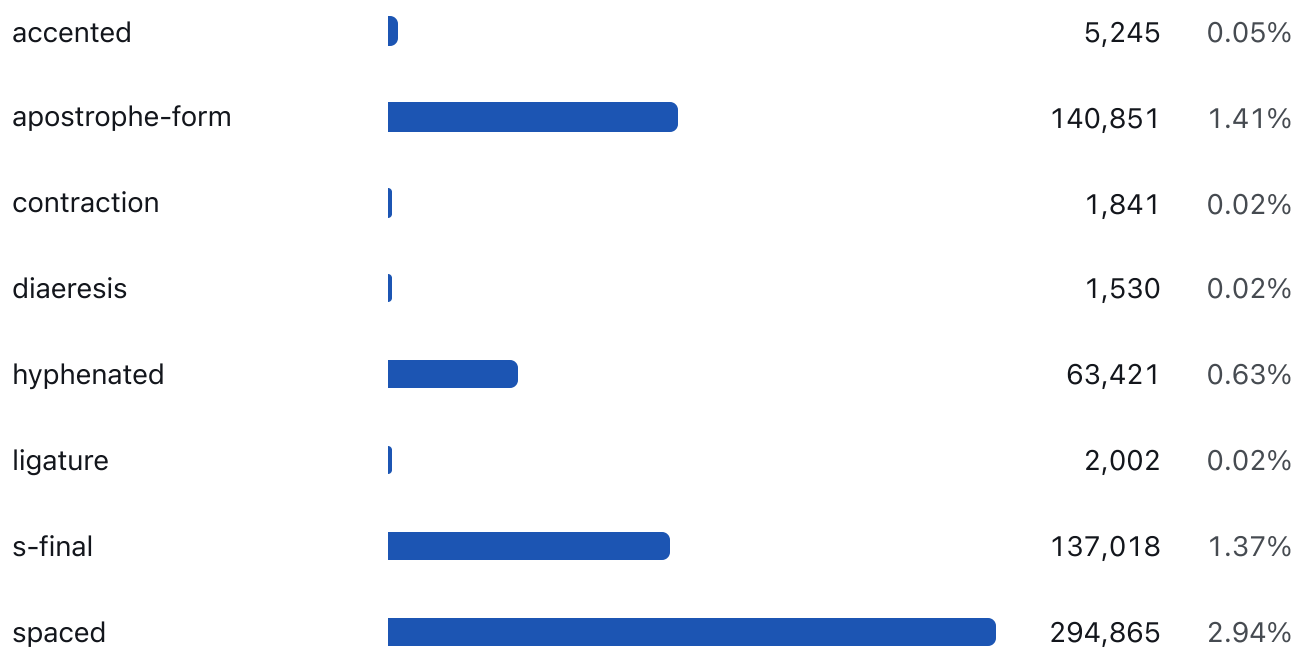
Group	Rows	10th percentile	Median	90th percentile	Longest
2+ families	417,864	6	9	13	45
1 family	672,104	7	10	15	100
corpus only	8,033,152	6	8	12	58
no evidence	602,165	7	11	15	63
multiword	294,865	9	14	21	79

Characters. All 50 characters of the charset occur. Of the accented letters and ligatures, the least used occurs in 39 rows and the most used in 2,794; the apostrophe occurs in 140,851 rows, the hyphen in 63,421, and the space in the 294,865 multiword rows. The site charts every character, by tier.

Classes. 9,538,019 rows (95.2%) are plain. The other 482,131 carry at least 1 class, in 48 combinations in all.

Rows per spelling class

A row can carry several classes; a row with none is plain.



Class flag	Rows	Share of all rows
`accented`	5,245	0.05%
`apostrophe-form`	140,851	1.41%
`contraction`	1,841	0.02%
`diaeresis`	1,530	0.02%
`hyphenated`	63,421	0.63%
`ligature`	2,002	0.02%
`s-final`	137,018	1.37%
`spaced`	294,865	2.94%

Apostrophes. Of the 140,851 rows with an apostrophe, 137,018 end in an apostrophe and s, 1,841 are contractions, and 2,033 are other forms.

Multiword terms. WordNet and Wiktionary both list 19,511 of the multiword rows, WordNet alone 43,972, and Wiktionary alone 231,382; every multiword row is listed by at least 1 of them, because they are its only sources.

Variants. 191,351 groups of 2 or more rows (401,577 rows, 4.0% of all) are equal once hyphens, spaces, apostrophes, and accents are removed. Most differ by an apostrophe (86,926 groups), a space (55,182), or a hyphen (24,553); the largest group has 7 members. Of the 174,500 pairs, a published list holds both members in 98,708, 1 in 71,530, and neither in 4,262.

Varies by	Groups
apostrophe	86,926
space	55,182
hyphen	24,553
hyphen, space	14,073
accent	4,191
hyphen, apostrophe	2,000
space, apostrophe	1,758
hyphen, space, apostrophe	1,218
apostrophe, accent	786
the same marks, placed differently	311

Lexical category

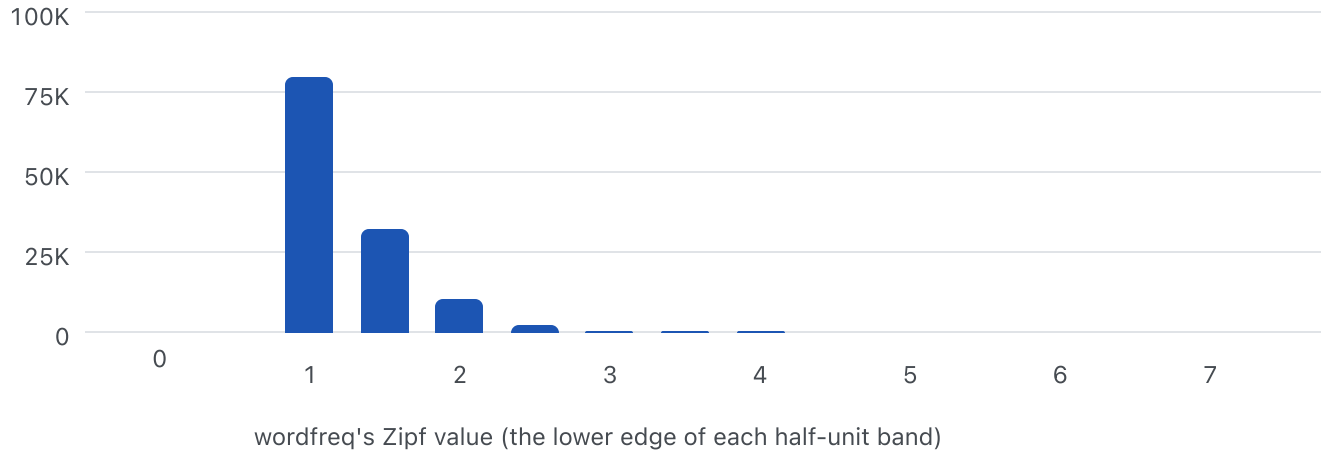
The universe records no part of speech, lemma, or inflection. WordNet lists 80,441 single-token rows and 63,483 multiword rows, and those rows could take a part of speech from WordNet's synsets; the rest have none. The plan for a later version adds a part-of-speech layer, lemma and inflection relations, and WordNet's categories. The morphology measures of the feature layer are among the withheld columns (§4).

Frequency

wordfreq scores few rows: 30.1% of the tier listed by 2 or more families, 6.7% of the tier listed by 1, and 1.7% of the corpus-only tier. Google Books holds more: 84.1%, 56.9%, and 99.9% of the same tiers. Where a value exists, the median Zipf value is 2.05 in the first tier, 1.60 in the second, and 1.35 in the corpus-only tier, and the median log10 count of Google Books matches is 3.95, 2.96, and 2.51.

wordfreq's Zipf value, by evidence tier

The rows wordfreq scores, by their Zipf value. Most rows have none, and the tier with no evidence has none by definition.



Shown: corpus only

Frequency rises with agreement. By the number of families that list a row, the median Google Books value climbs from 2.51 for rows no family lists to 5.89 for the 48,408 rows all 5 families list.

Families listing	Rows	Median Zipf	Median log10 Google Books matches
0	8,635,317	1.35	2.51
1	672,104	1.60	2.96
2	217,093	1.62	3.21
3	76,918	1.66	3.84
4	75,445	1.97	4.78
5	48,408	2.63	5.89

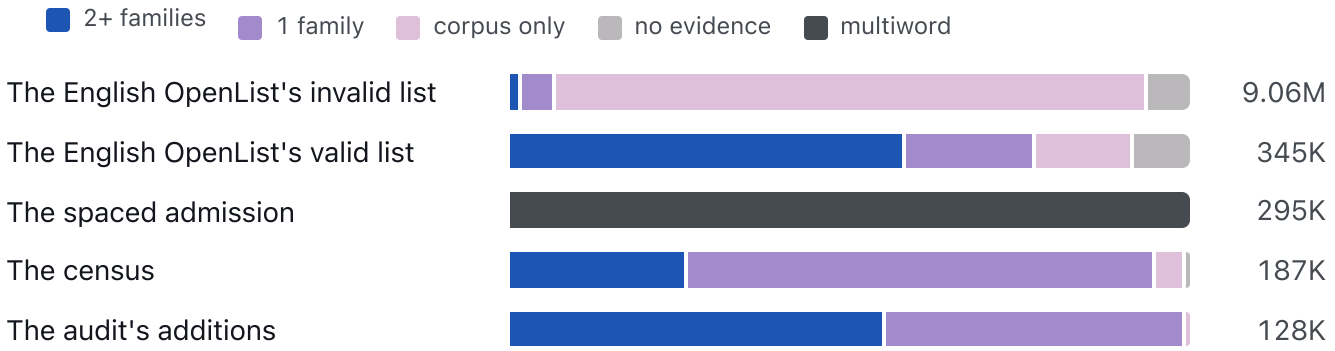
The span of years over which a string occurs in Google Books has a median of 252 years in the first tier, 160 in the second, and 181 in the corpus-only tier. It is a span in a corpus of scanned books, not a date of first use.

Provenance

Every row records its **intake layer** (the step of §5 that brought it in), its **intake kind** (obtained from a source, derived from pinned sources, or generated), and its first-seen **intake channels**, kept apart from the evidence for it. A channel records where a string was first seen and is never an attestation.

How each intake layer reaches each evidence tier

Each bar is 1 intake layer, its rows split by evidence tier; multiword rows are their own group.



Intake layer	Kind	2+ families	1 family	Corpus only	No evidence	Multiword
`c1_exceptions`	derived	48,789	131,219	7,240	18	0
`c4_additions`	derived	71,407	56,921	147	0	0
`eol_invalid`	generated	1,774	847	3,522	293,000	0
`eol_invalid`	obtained	93,495	418,348	7,973,125	280,720	0
`eol_valid`	generated	697	654	1,918	27,974	0
`eol_valid`	obtained	201,702	64,115	47,200	453	0
`spaced_c015`	derived	0	0	0	0	294,865

The EOL's invalid list contributes 9,064,831 rows, its valid list 344,713, the census 187,266, the audit 128,475, and the spaced admission 294,865. The largest channel, `eol:google_ngrams`, is the first-seen channel of 8,060,854 rows, of which a published family lists 5.0%. The site lists all 33 channels.

Intake channel	Rows	Listed by a family	In a corpus
`eol:google_ngrams`	8,060,854	5.0%	100.0%
`eol:wikipedia`	1,591,043	10.4%	84.2%
`eol:wiktionary`	635,145	80.2%	78.1%
`eol:synthetic-engine`	330,386	1.2%	2.6%
`eol:phase1-unrecorded`	303,604	2.7%	94.3%
`spaced:wiktionary_english`	250,893	100.0%	0.0%
`eol:channel-withheld`	210,109	83.5%	97.0%
`eol:opensource`	171,255	92.6%	81.4%
`audit:wiktionary_english`	104,848	100.0%	80.0%
`audit:scowl`	90,637	100.0%	80.8%
`census:hunspell_en_GB`	86,480	100.0%	16.2%
`census:scowl`	79,418	100.0%	10.3%

The synthetic engine. 330,386 rows were made by the EOL's synthetic engine. Its code does not survive, and its own per-row record of the method is empty, but its author describes the methods: affixes applied without a part-of-speech filter to the list as it then stood (plurals, past tenses, gerunds, agent nouns, comparatives and superlatives, and prefixed forms); chemical names varied by substituent, and short chains of root, infix, and suffix, made by substitution over a hand-written list of substituents and roots, with no tool that derives a name from a formula; medical and scientific prefixes; and Greek plurals. It gave each base word a domain label, and it used NLTK, lemminflect, and spaCy. Each row carries the method inferred from its form. A published family lists 3,972 of them (1.2%).

Engine, as inferred	Rows	Listed by a family
affixation/plural	149,416	3,279
affixation/verb-form	117,315	446
affixation/superlative	30,682	76
affixation/comparative-or-agent	29,940	121
chemical-variation/substituent-prefix	903	0
greek-plural/-ses	812	23
chemical-variation/root-chain	691	4
medical-scientific-prefixing/domain-prefix	338	12
greek-plural/-des	142	1
unresolved/none	111	5
affixation/general-prefix	36	5

Provenance classes. Over the published fields, 8,920,645 rows have both a recorded channel and containment in a list or a corpus, 585,792 a channel only, 497,340 containment only, and 16,373 neither.

The EOL since its pin. The EOL's last published list (revision `c707c187b2c3`) holds 345,162 entries: the 345,103 of its list at the pin and 59 added after it. The universe's `eol-valid` config, which is the EOL's public dataset, holds 344,713; the license rules hold back the other 390.

7. The sources compared

This section is the one no single list can write, because none measures the same strings against the others. It covers the single-token rows, where all 8 lists are measured.

Methods. The main pair measure is directed containment: of the strings list A lists that list B could hold, the share B lists. It is computed over the jointly eligible rows, with A's rows that B cannot hold reported beside it as a difference of scope, not of judgment. The Jaccard index appears as a secondary measure only, because it cannot exceed the smaller list's size over the

larger's. Agreement is counted as k of m: the families that list a row out of the families that could. The report makes no significance test, because at this size every test is significant and the null hypothesis is false by construction, and it makes no estimate of how many English strings no list holds: the lists derive from one another, listing tracks frequency, and the target is open-ended.

Sizes and blind classes

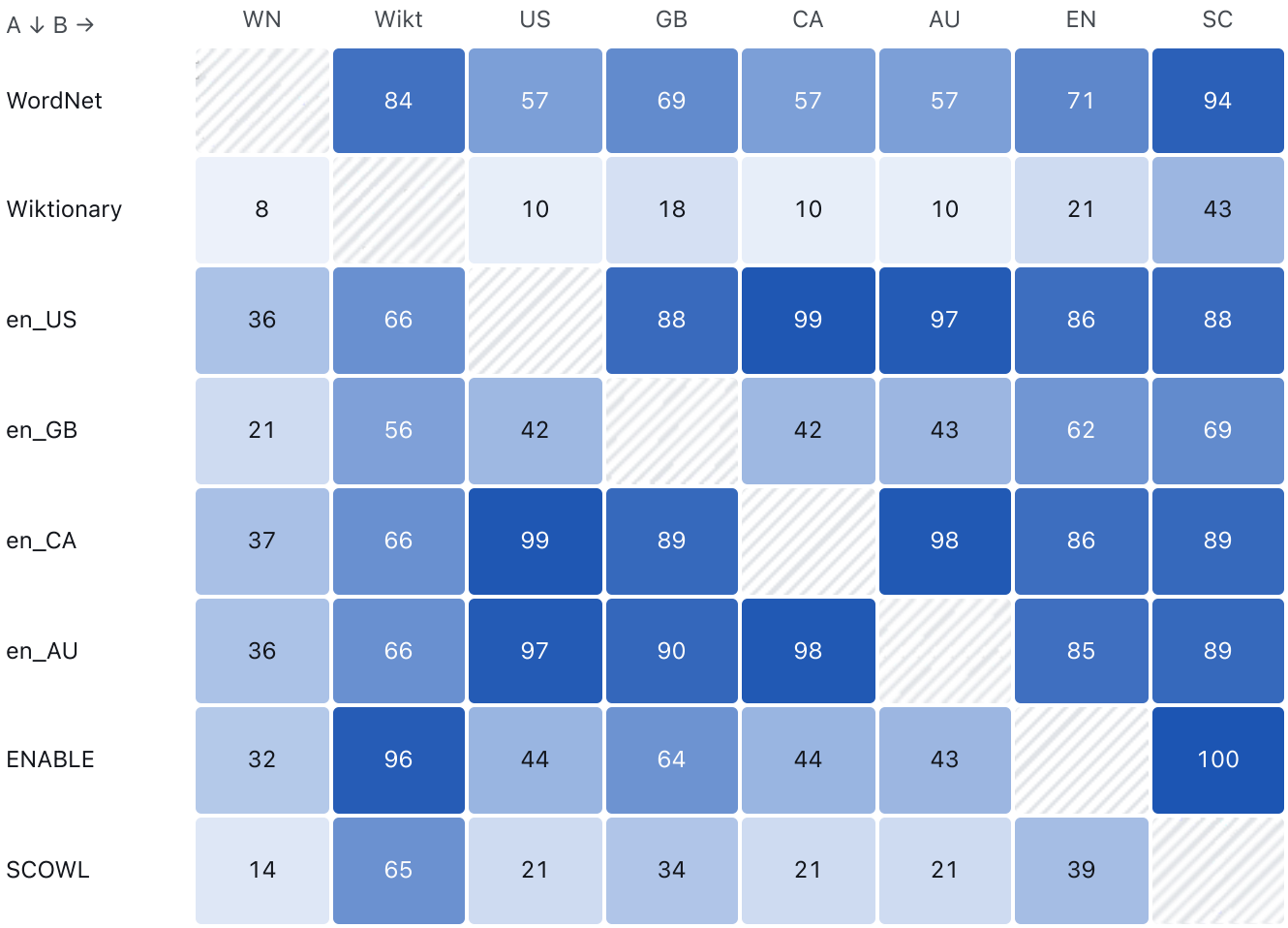
List	Listed	Cannot hold	Listed, share of single-token rows
WordNet	80,441	8,147	0.83%
Wiktionary (English)	831,973	0	8.55%
hunspell en_US	121,252	61,747	1.25%
hunspell en_GB	262,413	1,940	2.70%
hunspell en_CA	120,947	61,747	1.24%
hunspell en_AU	121,490	61,747	1.25%
ENABLE	172,823	187,266	1.78%
SCOWL	518,438	55,759	5.33%

Wiktionary lists the most single-token rows, 831,973, and is blind to no class. WordNet lists the fewest, 80,441. ENABLE cannot hold 187,266 rows, every row with an exception class, and each hunspell dictionary but the British cannot hold 61,747.

Containment

How far each list contains the others

Of the rows list A lists that list B can hold, the share B lists. Read across a row; the map is not symmetric, so nesting shows as a bright cell beside a dim one. Columns: WN WordNet, Wikt Wiktionary, US, GB, CA, and AU the hunspell dictionaries, EN ENABLE, SC SCOWL.



Shown: containment

A, listed by B →	WN	Wikt	US	GB	CA	AU	EN	SC
WN	—	84%	57%	69%	57%	57%	71%	94%
Wikt	8%	—	10%	18%	10%	10%	21%	43%
US	36%	66%	—	88%	99%	97%	86%	88%
GB	21%	56%	42%	—	42%	43%	62%	69%
CA	37%	66%	99%	89%	—	98%	86%	89%
AU	36%	66%	97%	90%	98%	—	85%	89%
EN	32%	96%	44%	64%	44%	43%	—	100%
SC	14%	65%	21%	34%	21%	21%	39%	—

The heat map's asymmetry shows nesting. ENABLE's rows that SCOWL can hold are 99.7% listed by SCOWL, and SCOWL's that ENABLE can hold are 39.2% listed by ENABLE: ENABLE is

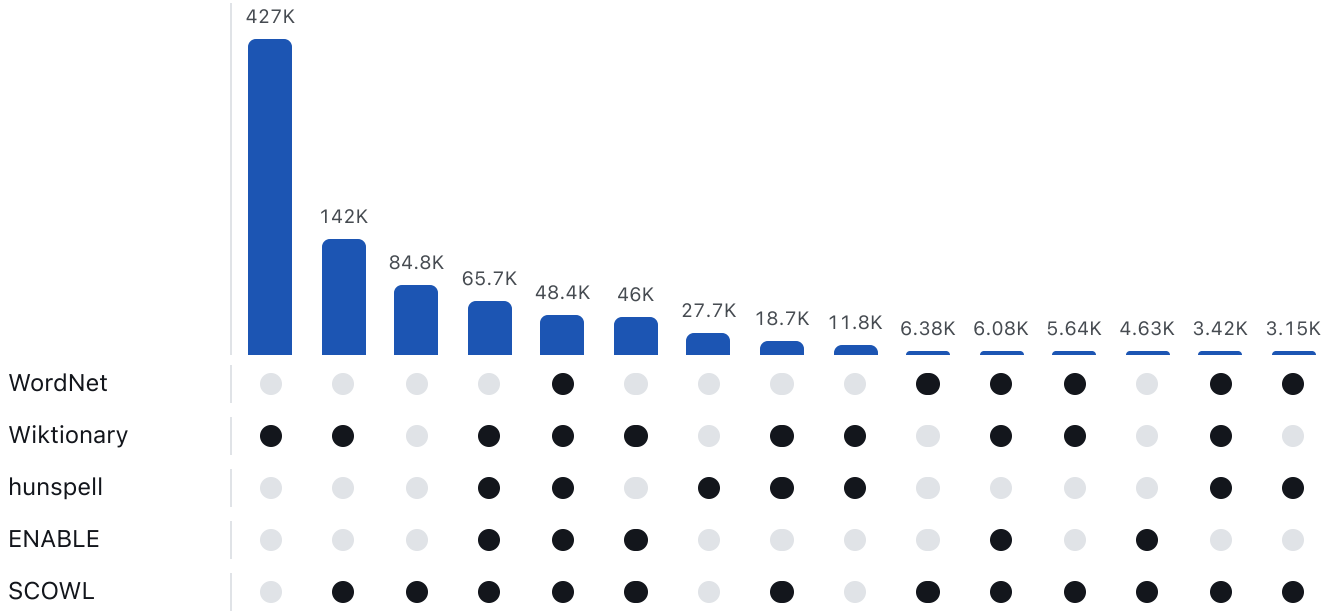
nearly a subset. The American hunspell dictionary's rows are 98.9% listed by the Canadian. WordNet's rows are 94.5% listed by SCOWL and 83.8% by Wiktionary, while Wiktionary's are 8.1% listed by WordNet and 43.1% by SCOWL: Wiktionary holds much the others do not.

Combinations and what each family holds alone

Over the 9,538,019 rows every family can hold, the rows fall in 32 combinations of families. 8,628,059 (90.5%) are listed by no family. Of the rest, 426,798 are listed by Wiktionary alone, and 48,408 (5.3%) by all 5.

Which families list the rows

Over the 9,538,019 rows all 5 families can hold; the 15 largest combinations. A filled dot marks a family that lists the rows.



Families that list the row	Rows
none	8,628,059
Wiktionary	426,798
Wiktionary, SCOWL	141,928
SCOWL	84,810
Wiktionary, hunspell, ENABLE, SCOWL	65,712
WordNet, Wiktionary, hunspell, ENABLE, SCOWL	48,408
Wiktionary, ENABLE, SCOWL	46,033
hunspell	27,682
Wiktionary, hunspell, SCOWL	18,713
Wiktionary, hunspell	11,824

Families that list the row	Rows
WordNet, SCOWL	6,381
WordNet, Wiktionary, ENABLE, SCOWL	6,076

Over all single-token rows, the rows only 1 family lists are:

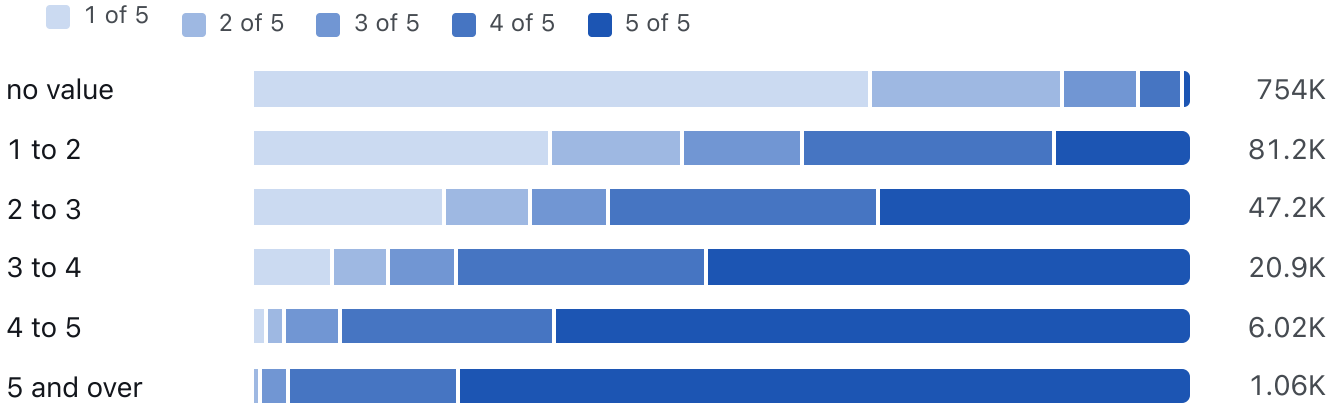
Family	Rows only it lists	Share of listed rows
WordNet	1,522	0.1%
Wiktionary	476,575	43.7%
hunspell	72,784	6.7%
ENABLE	78	0.0%
SCOWL	121,145	11.1%

Agreement

Among the rows all 5 families can hold, 909,960 are listed by at least 1. Of those, 59.4% are listed by exactly 1 family and 5.3% by all 5.

How many families list a row, by frequency

The listed rows all 5 families can hold, by the number of families that list them, in each band of wordfreq's Zipf value.



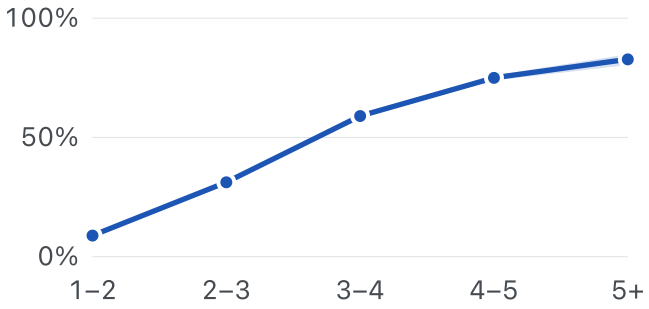
Families that can hold (m)	k = 0	k = 1	k = 2	k = 3	k = 4	k = 5
1	14	1,926				
2		212	7			
3	2,418	50,968	4,810	1,392		
4	4,826	78,113	42,376	190	14	
5	8,628,059	540,885	169,900	75,336	75,431	48,408

Coverage by frequency

Coverage by frequency

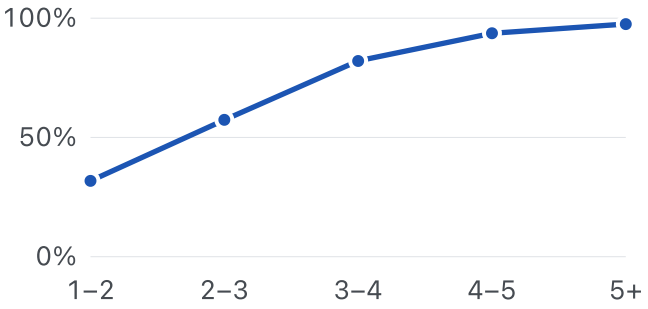
For each band of wordfreq's Zipf value, the share of the rows a list can hold that it lists. Most rows have no Zipf value; that band is stated under each chart.

WordNet



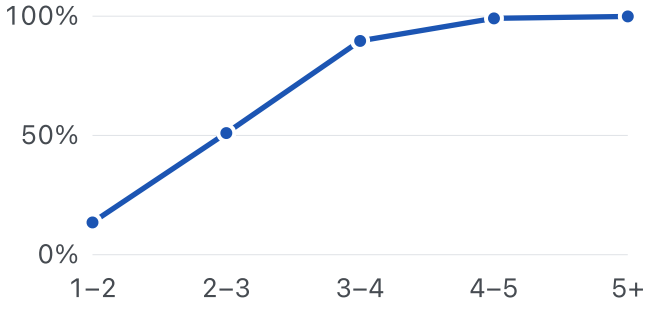
No Zipf value: 0.3% listed

Wiktionary (English)



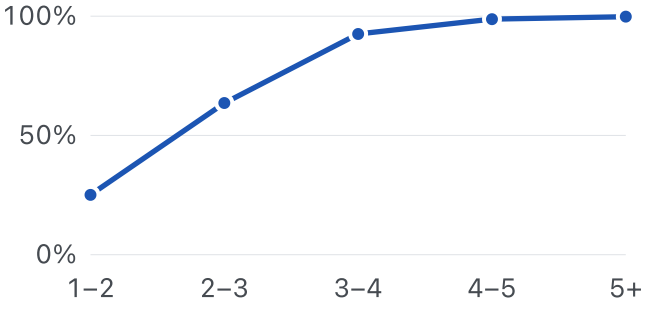
No Zipf value: 7.5% listed

hunspell en_US



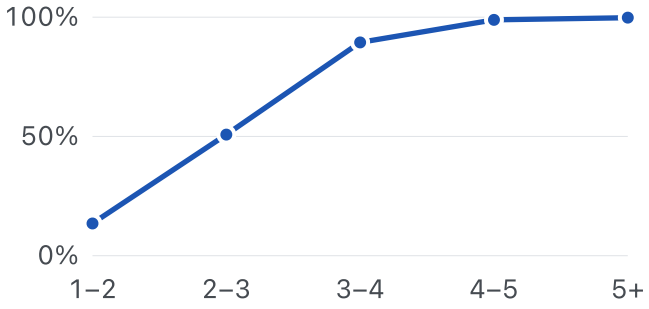
No Zipf value: 0.4% listed

hunspell en_GB



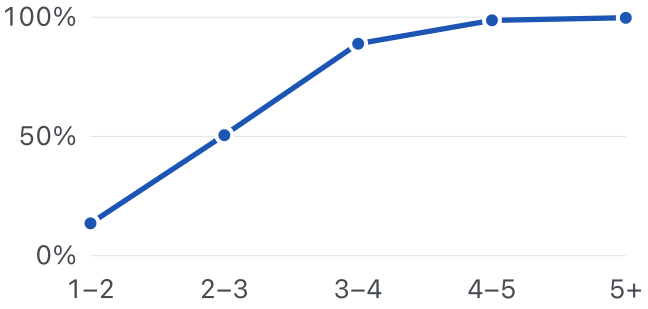
No Zipf value: 1.5% listed

hunspell en_CA

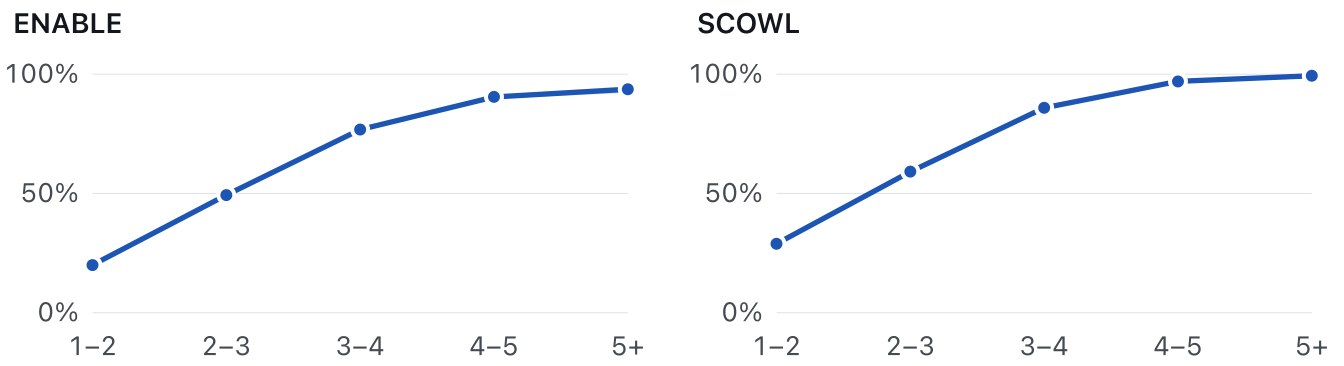


No Zipf value: 0.4% listed

hunspell en_AU



No Zipf value: 0.4% listed



No Zipf value: 0.9% listed

No Zipf value: 4.2% listed

Shown: shares by strings

For each band of wordfreq's Zipf value, the share of the rows a list could hold that it lists. Rows wordfreq does not score form their own band, which holds most rows. Coverage rises with frequency for every list. In the band of Zipf 5 and over, hunspell's American dictionary lists 99.9% and WordNet 82.7%; weighted by tokens, WordNet's share there falls to 57.4%: WordNet holds only nouns, verbs, adjectives, and adverbs, and the most frequent strings include function words.

Zipf band	Single-token rows
no value	9,421,774
1 to 2	210,302
2 to 3	64,381
3 to 4	21,668
4 to 5	6,071
5 and over	1,089

By strings:

List	no value	1 to 2	2 to 3	3 to 4	4 to 5	5 and over
WordNet	0%	9%	31%	59%	75%	83%
Wiktionary (English)	7%	32%	57%	82%	94%	98%
hunspell en_US	0%	14%	51%	90%	99%	100%
hunspell en_GB	2%	25%	64%	93%	99%	100%
hunspell en_CA	0%	14%	51%	89%	99%	100%
hunspell en_AU	0%	14%	51%	89%	99%	100%
ENABLE	1%	20%	49%	77%	90%	94%
SCOWL	4%	29%	59%	86%	97%	99%

Weighted by tokens:

List	1 to 2	2 to 3	3 to 4	4 to 5	5 and over
WordNet	11%	38%	62%	77%	57%

List	1 to 2	2 to 3	3 to 4	4 to 5	5 and over
Wiktionary (English)	36%	63%	85%	95%	99%
hunspell en_US	18%	61%	93%	99%	100%
hunspell en_GB	31%	72%	94%	99%	100%
hunspell en_CA	18%	60%	93%	99%	100%
hunspell en_AU	18%	60%	92%	99%	100%
ENABLE	24%	56%	80%	92%	93%
SCOWL	33%	66%	89%	98%	100%

The varieties, the core, and the tails

American and British spellings. 14,308 rows are in the American hunspell dictionary and not the British, 146,322 in the British and not the American, and 106,944 in both. The 4 hunspell dictionaries all list 106,399 rows; of the rows at least 1 of them lists, 60.3% are listed by some and not all.

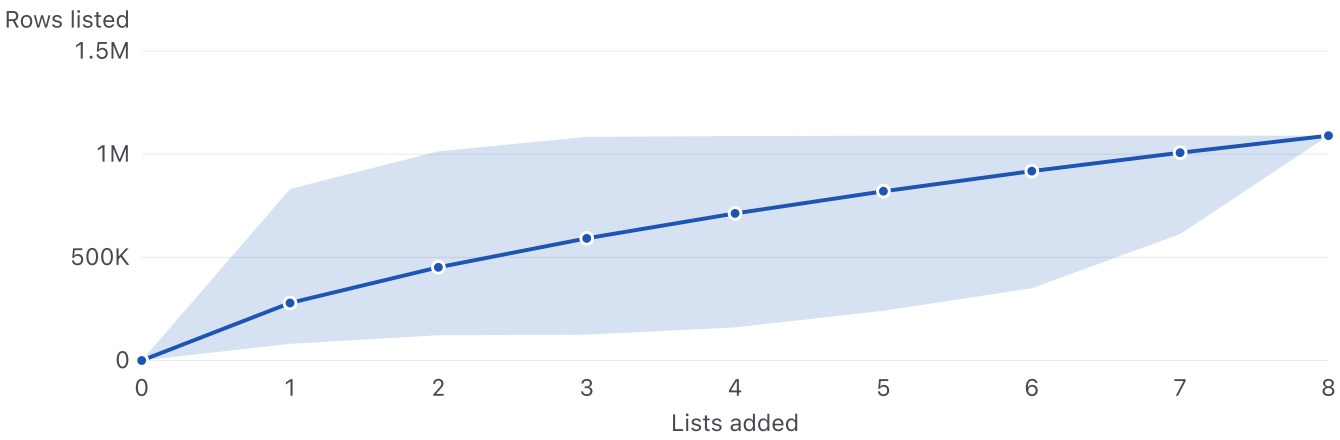
The consensus core. 417,864 rows (4.3%) are listed by 2 or more families. The EOL's valid list contributes 202,399 of them and its invalid list 95,269, strings the EOL ruled out that 2 or more independent families list.

The tails. 378 rows with a Zipf value of 3 or more are listed by no family. In the other direction, 347,609 listed rows (31.9% of the listed rows) are in neither corpus.

What each list adds

How the union grows as lists are added

For each number of lists, the mean union over every choice of that many lists, and its smallest and largest.



Without a given list, the rows only it lists would be unlisted (§7). As lists are added in every order, the union grows from a mean of 278,722 rows for 1 list to 1,089,968 for all 8; the order matters, as the range at each step shows.

Lists	Mean union	Smallest	Largest
1	278,722	80,441	831,973
2	451,746	122,249	1,013,749
3	592,049	124,889	1,084,450
4	713,157	160,092	1,088,351
5	820,725	240,688	1,089,887
6	918,139	349,973	1,089,965
7	1,007,439	613,393	1,089,968
8	1,089,968	1,089,968	1,089,968

8. Limitations

- **Most rows rest on a single corpus.** 8,024,536 single-token rows (80.08%) were inherited from the EOL, are listed by none of the 8 published panel sets, and are Google Books 1-gram types. For 7,900,248 of them, Google Books is the only published containment, and it is a corpus built from scanned text, which carries recognition errors. In all, 8,761,768 rows are Google Books types.
- **602,165 single-token rows are listed by no published panel set and are in neither corpus.** What supports them is their intake channel, where a published one is recorded.
- **16,373 rows (0.16%) have no published intake channel and no measured containment.** Nothing published says how the string entered the pool, and no published panel set or corpus holds it.
- **330,386 rows (3.30%) were made by the EOL's synthetic engine.** Its code does not survive, and its own per-row record of the method is empty, but its author describes the methods: affixes applied without a part-of-speech filter to the list as it then stood (plurals, past tenses, gerunds, agent nouns, comparatives and superlatives, and prefixed forms); chemical names varied by substituent, and short chains of root, infix, and suffix, made by substitution over a hand-written list of substituents and roots, with no tool that derives a name from a formula; medical and scientific prefixes; and Greek plurals. It gave each base word a domain label, which `eol_synthetic_category` carries where it survives, and it used NLTK, lemminflect, and spaCy. `model_or_engine` marks these rows `engine_generated`, and each row's method is inferred from its form (§8).
- **137,485 rows of the EOL's valid list were decided with a language model:** 131,014 by the EOL's verification pipeline, which used one, and 6,471 by a model's verdict alone.

`model_or_engine` marks them `model_pipeline` and `model_alone` (§11). The EOL's invalid list is not marked.

- **The withholding changes what the panel columns count.** The build measured 13 panel sets, and 5 of them are withheld (§4). `panel_sets`, `panel_count`, and `provenance_class` count the 8 published sets alone, so a row with `panel_count` 0 may still be listed by a withheld list.
- **274,471 rows of the shippable projection are not in this version.** Once the withheld lists and channels are taken away, nothing published supports them, so they are held back in the private record (§4).
- **A channel may be withheld for license reasons.** 210,109 rows carry `eol:channel-withheld` in `intake_channels`: every channel recorded for them is withheld, because it names a withheld list or a source whose terms do not allow publishing the label.
- **12 of the feature layer's columns are withheld,** because each was computed against a reference that holds a withheld list, or by a character model fitted on strings the withheld lists help choose (§4).
- **The `sources` column has 2 meanings:** a per-source containment list on the CEOL's own rows, and the dataset-level tag `eol_universe@c1139698` on inherited rows, whose per-source listing is the evidence columns'.
- **Null and NaN mean different things** (§3), and the corpora do not measure multiword rows.
- **`eol` is a fact at a single revision of the EOL, `c1139698`.** The EOL's last published list adds 59 entries made after that revision; they are rows of this dataset with `eol` false.

What the universe does not record: part of speech; lemma and inflection relations; definitions and senses; pronunciation, syllables, etymology, and a first attestation year; variety and register labels (the 2 variety columns are proxies); a frequency drawn from many kinds of text for most rows (wordfreq scores few; the Google Books counts cover most single-token rows, but from scanned books alone); and validity, by design.

9. Uses and out-of-scope uses

Studies that build lists of real or invented words each check existence against a reference of their own, and none stores a per-item, per-reference result. The universe stores one. It supports, with these caveats:

1. **A "listed by 0 of N lists" screen for invented stimuli,** never a "not a word" verdict. N is the published lists of the version, here 8.

2. **Auditing pseudoword generators:** finding generated strings that a list holds, or that are variants of a listed string (§6).
3. **Graded existence per list,** with "not measured" kept apart from "not listed".
4. **A reproducible frame for choosing items** in rating and norming studies: a stated filter over a pinned version, with its code.
5. **Control of regional spellings** through the variety proxies.
6. **Frequency matching in the very low range,** with missing values explicit.
7. **A dated exclusion list** for benchmarks of novel words, re-run at each version.
8. **Stratified pools** for vocabulary tests.

How to check a made-up string. Look it up on the site or in the `universe` config; read its listing in each list, remembering that -1 means the list cannot hold it; and treat a listing as a listing, not as use, and absence from every list as absence from the lists, not from text. A multiword string is checked against 2 lists only.

Out of scope: calling a string a word or not a word; using the universe as a dictionary; proper names; and any claim about what a model was trained on.

10. Maintenance and versions

A released version never changes; a correction comes as a new version with a change log (`docs/UPKEEP.md`). A row leaves a public version only under the rules of that policy, and a version may be withdrawn, never edited. A request to withhold, correct, or restore a source or a row is acknowledged within 7 days and acted on within 14 days of its verification, and each version's change log records every withholding, as counts. Each version records the commit of the code that built it, and a checkpoint measures every source again and builds a new pin and version.

11. Responsible use and disclosures

- **The study's material stays private.** The project published the design of a study of words in use as a proposal; it has not been run. Its development ledger, synthetic controls, probe batteries, pilot items, and every label and verdict stay private, and the build asserts that none of the synthetic controls is in any file.
- **Language models.** No language model decided any row the CEOL added. The EOL decided some rows of its valid list with a language model, Google Gemini 3 Flash Preview by its per-

entry record, and `model_or_engine` marks those rows, so a list can keep them or leave them out.

- **Offensive and personal content.** The universe holds strings its sources hold, and some are offensive. This version does not measure their share: the analysis plan's measure needs an offensive-term lexicon whose license allows the use, which the project has not chosen. No string is printed in this report.
- **Contamination.** The release's date, file hashes, and canary are recorded, so that a later question of whether a model saw the dataset can be dated. The project's own pages mention strings; they are mentions, not uses.

12. Availability and reproducibility

The dataset is `ryanjosephkamp/ceol-universe` on the Hugging Face hub, at the tag of its version, with a Croissant file carrying its responsible-AI fields. The code and every committed record are at https://github.com/ryanjosephkamp/ceol-universe. The release rebuilds from its pinned layers with `python3 -m research.release.build_release --version v1.0`, the descriptive analysis with `python3 -m research.analysis.descriptive --version v1.0`, and this report with `python3 -m research.release.report.paper --render`. The analysis is deterministic: its check rebuilds `analysis.json` and compares it byte for byte. The pinned layers themselves are private, because they hold the withheld lists; anyone who holds the same versions of the sources can rebuild them from the archived inputs the records name.

Appendix A, the comparison over all 13 lists of the private universe, waits for the license review's confirmation that the terms under which the project holds each withheld list allow publishing aggregate counts.

Appendix B. The columns

The `universe` config carries 41 columns and the `features` config 11.

The key

Column	Type	Meaning	Null
`id`	string	the first 16 hexadecimal digits of the sha256 of the word's UTF-8 bytes; a function of the string alone, so the same string carries the same id in every version	never null

Column	Type	Meaning	Null
`word`	string	the candidate string, the output of the grammar: lower case, the 50-character charset, spaces single and interior; the join key of every layer	never null

The pin's columns

Column	Type	Meaning	Null
`eol`	bool	in the EOL's valid list at the pinned revision c1139698: a historical fact at a pin, carried as a covariate, never a CEOL verdict	never null
`eol_universe`	bool	in the EOL's pinned candidate universe (its valid or its invalid list); false for the CEOL's own 3 layers	never null
`classes`	string	the grammar's non-exclusive class flags joined by ';' (hyphenated, accented, diaeresis, ligature, apostrophe-form, contraction, s-final, spaced); the empty string on a plain-letter string	never null
`sources`	string	provenance at intake: for the CEOL's own layers the published sources whose derived sets contain the string, joined by ';', a withheld list's label removed (the empty string where each such source is withheld); for the strings inherited from the EOL the dataset-level tag eol_universe@c1139698 (their per-source listedness is the evidence columns')	never null
`intake_date`	string	the date the row entered a CEOL artifact (an ISO date): August 17, 2026 the EOL universe at the pinned revision, August 20, 2026 the census's exception candidates and the audit's additions, August 26, 2026 the spaced admission's multiword candidates	never null
`c4_addition`	bool	a row of the audit's additions to the single-token pin	never null
`is_restricted`	bool	attested only by NWL2023 or CSW21; false on every row of this dataset, because those rows never ship	never null
`token_count`	int32	the number of space-separated tokens	never null
`c015_addition`	bool	a row of the multiword candidates of the spaced admission	never null
`wn_instance`	bool	WordNet instance-synset membership, a properness signal recorded and never filtered on	null unless WordNet attests the row; measured on the spaced layer only; null means not measured, never false

Column	Type	Meaning	Null
`wn_capitalized`	bool	WordNet data-file capitalization, a properness signal	null unless WordNet attests the row; measured on the spaced layer only
`wk_proper_noun`	bool	a Wiktionary Proper-noun heading, a properness signal	null unless Wiktionary attests the row; measured on the spaced layer only
`wk_flags`	string	Wiktionary's category flags joined by ';' (idiomatic, phrasal-verb, phrasebook, proverb); may be the empty string	null unless Wiktionary attests the row; measured on the spaced layer only

The release layer

Column	Type	Meaning	Null
`class_explicit`	string	the pin's classes with the value 'plain' in place of the empty string; otherwise the flags unchanged	never null
`panel_sets`	string	the published panel sets that list the row, joined by ';' in the order wordnet, wiktionary_english, hunspell_en_US, hunspell_en_GB, hunspell_en_CA, hunspell_en_AU, enable1, scowl; on a single-token row a binary set lists the row where its evidence column is 1 and SCOWL where scowl_tier is above 0; on a spaced row the sets are the pin's published sources; the empty string where none lists the row. A withheld list is never named	never null
`variety_gb_not_us`	bool	true where hunspell_en_GB is 1 and hunspell_en_US is 0: a proxy for a variety difference between the 2 dictionaries, never a variety label	null where either hunspell column is -1 or null
`variety_us_not_gb`	bool	true where hunspell_en_US is 1 and hunspell_en_GB is 0: the reverse proxy	null where either hunspell column is -1 or null

The intake-provenance layer

Column	Type	Meaning	Null
`intake_layer`	string	the layer the row entered by: eol_valid, eol_invalid, c1_exceptions, c4_additions, spaced_c015	never null
`intake_kind`	string	obtained (an inherited string taken in through a channel), derived (a row of a CEOL layer derived from pinned sources), or generated (made by the EOL's lost synthetic engine)	never null

Column	Type	Meaning	Null
`intake_channels`	string	the first-seen channel labels joined by ';': the EOL's discovery log with its tiers removed, and the CEOL layers' own namespaces census:, audit:, spaced:, ledger:; a record of where a string was first seen, never an attestation. A label naming a withheld list is removed, and so is a channel withheld for license reasons; a row whose every label was withheld carries the single label eol:channel-withheld	never null
`eol_phase1_tier`	string	the tier suffix the EOL's labels on the row share: valid, unlikely, or none (mixed where the labels disagree)	never null
`eol_record`	string	which EOL dictionary holds the row's entry: valid-dict, invalid-dict, none	never null
`eol_synthetic_category`	string	the category recorded on a generated string (a domain label, or the EOL validation step's judgment)	null on every row that is not generated
`eol_synthetic_engine_inferred`	string	the engine and rule inferred from a generated string's form (affixation, chemical-variation, greek-plural, medical-scientific-prefixing, unresolved); an inference, so named	null on every row that is not generated
`panel_count`	int16	the number of the 8 published panel sets that list the row (the number of names in panel_sets, asserted equal on every row); a withheld list's listing is not counted	never null
`gb_type`	bool	the string is a Google Books English 1-gram type in the CEOL re-stream	null on spaced rows, which the corpora do not measure
`wordfreq_scored`	bool	wordfreq 3.1.1 scores the string	null on spaced rows, which the corpora do not measure
`wikt_title_any`	bool	some language's Wiktionary has a page titled with the string: a screen, not containment	null on spaced rows
`provenance_class`	string	exactly 1 of channel_and_containment, channel_only, containment_only, neither, over the published fields: whether the row has a published intake channel (a label other than eol:phase1-unrecorded and eol:channel-withheld) and whether a published panel set or a corpus contains it	never null

The evidence matrix

Column	Type	Meaning	Null
`wordnet`	int8	listed by WordNet 3.0's lemmas, re-derived under the CEOL grammar: 1 listed, 0 not listed, -1 structurally blind to the row's class	null on spaced rows, which the single-token sets were not measured over
`wiktionary_english`	int8	listed by Wiktionary pages with an English section (the August 4, 2026 dump): 1 listed, 0 not listed, -1 structurally blind to the row's class	null on spaced rows, which the single-token sets were not measured over
`hunspell_en_US`	int8	listed by the LibreOffice hunspell en_US dictionary, expanded: 1 listed, 0 not listed, -1 structurally blind to the row's class	null on spaced rows, which the single-token sets were not measured over
`hunspell_en_GB`	int8	listed by the LibreOffice hunspell en_GB dictionary, expanded: 1 listed, 0 not listed, -1 structurally blind to the row's class	null on spaced rows, which the single-token sets were not measured over
`hunspell_en_CA`	int8	listed by the LibreOffice hunspell en_CA dictionary, expanded: 1 listed, 0 not listed, -1 structurally blind to the row's class	null on spaced rows, which the single-token sets were not measured over
`hunspell_en_AU`	int8	listed by the LibreOffice hunspell en_AU dictionary, expanded: 1 listed, 0 not listed, -1 structurally blind to the row's class	null on spaced rows, which the single-token sets were not measured over
`enable1`	int8	listed by the ENABLE word list: 1 listed, 0 not listed, -1 structurally blind to the row's class	null on spaced rows, which the single-token sets were not measured over
`scowl_tier`	int16	the SCOWL size tier (10 to 95) of the smallest SCOWL list holding the string; 0 not listed; -1 not applicable to the row's class	null on spaced rows
`wikt_title_screen`	int8	1 where some language's Wiktionary has a page titled with the string, else 0: a screen, never a detector	null on spaced rows

Model and engine

Column	Type	Meaning	Null
`model_or_engine`	string	engine_generated on a row the EOL's synthetic engine made (intake_kind generated); on a row of the EOL's valid list, model_pipeline where the EOL listed it through its verification pipeline, which used a language model, and model_alone where a language model's verdict is its only basis; none on every other row. A record of how the EOL decided, never a verdict of this dataset's; the EOL's invalid list is not marked	never null

The feature layer

Column	Type	Meaning	Null
`id`	string	the first 16 hexadecimal digits of the sha256 of the word's UTF-8 bytes; a function of the string alone, so the same string carries the same id in every version	never null
`word`	string	the candidate string, the output of the grammar: lower case, the 50-character charset, spaces single and interior; the join key of every layer	never null
`gb_year_span`	int16	the years from the first to the last in which the string occurs in the Google Books English 1-grams (the 2020 release), inclusive	0 where the string is not a 1-gram type
`gb_span_fill`	float	the share of the span's years in which the string occurs	NaN where the string is not a 1-gram type
`gb_tokens_per_volume`	float	the total match count over the total volume count	NaN where no volume carries the string
`gb_volumes_per_year`	float	the total volume count over the years in which the string occurs	NaN where the string is not a 1-gram type
`gb_log10_match`	float	log10 of the total match count	NaN where the string is not a 1-gram type
`gb_log10_volume`	float	log10 of the total volume count	NaN where no volume carries the string
`zipf_log10_rank`	float	log10 of the string's frequency rank among the aggregate's types	NaN where the string is not a 1-gram type
`zipf_resid`	float	log10 of the match count minus the Zipf-Mandelbrot fit's prediction at that rank	NaN where the string is not a 1-gram type
`wordfreq_zipf`	float	the wordfreq 3.1.1 Zipf frequency (English) where wordfreq's list holds the string	NaN elsewhere, and on every hyphenated string (wordfreq is hyphen-blind)

Glossary

- **Candidate:** a string that passes the spelling grammar; every row is one.
- **Listed:** a list, at its pinned version, includes the string.
- **Cannot hold (blind):** a list structurally excludes strings of the row's class.
- **Family:** the lists that derive from 1 upstream; hunspell's 4 dictionaries are 1 family.
- **Evidence tier:** a single-token row's count of listing families, with "corpus only" and "no evidence" for rows no family lists.
- **Intake channel:** where a string was first seen; never an attestation.
- **Withheld list:** a list whose columns and labels the public version drops.
- **Restricted row:** a row only the withheld proprietary lists attest; it is in no version.

- **Held-back row:** a shippable row with no published evidence once the withholding is applied.

References

1. North American Scrabble Players Association, "SOWPODS," NASPA Wiki. https://www.scrabbleplayers.org/w/SOWPODS. Read October 26, 2026.
2. Wordnik, "Wordnik For Game Developers," developer.wordnik.com. https://developer.wordnik.com/gameset. Read October 26, 2026.
3. Princeton University, "WordNet: A Lexical Database for English," home page. https://wordnet.princeton.edu/. Read October 26, 2026, through the Internet Archive capture of October 6, 2026 (the live site refused the request).
4. Princeton University, "wnstats(7WN): WordNet 3.0 database statistics," WordNet documentation. https://wordnet.princeton.edu/documentation/wnstats7wn. Read October 26, 2026, through the Internet Archive capture of October 3, 2026.
5. Princeton University, "License and Commercial Use of WordNet," WordNet documentation. https://wordnet.princeton.edu/license-and-commercial-use. Read October 26, 2026, through the Internet Archive capture of October 4, 2026.
6. Princeton University, "wndb(5WN): WordNet database files," WordNet documentation. https://wordnet.princeton.edu/documentation/wndb5wn. Read October 26, 2026, through the Internet Archive capture of May 10, 2026.
7. Open English WordNet team, "english-wordnet: README.md, LICENSE.md, and FORMAT.md," GitHub repository globalwordnet/english-wordnet, branch main. https://github.com/globalwordnet/english-wordnet. Read October 26, 2026.
8. Open English WordNet team, "Downloads," en-word.net. https://en-word.net/downloads. Read October 26, 2026.
9. Wiktionary contributors, "Wiktionary:Statistics," English Wiktionary, revision 92513152, table generated from the October 1, 2026 dump. https://en.wiktionary.org/wiki/Wiktionary:Statistics. Read October 26, 2026.
10. T. Ylonen, "Wiktextract: Wiktionary as Machine-Readable Structured Data," in Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), Marseille, pp. 1317–1325, 2022. https://aclanthology.org/2022.lrec-1.140/.
11. Wikimedia Foundation, "Terms of Use," effective June 7, 2023, section 7, "Licensing of Content". https://foundation.wikimedia.org/wiki/Policy:Terms_of_Use. Read October 26, 2026.
12. kaikki.org, "English machine-readable dictionary; Raw data downloads; dictionary index," kaikki.org. https://kaikki.org/dictionary/English/index.html,

https://kaikki.org/dictionary/rawdata.html, and https://kaikki.org/dictionary/index.html. Read October 26, 2026.

13. G. Sérasset, "Dbnary – Wiktionary as Linguistic Linked Open Data," home page. http://kaiko.getalp.org/about-dbnary/. Read October 26, 2026, through the Internet Archive capture of October 11, 2026 (the live site did not respond).
14. G. Sérasset, "DBnary - Wiktionary Lexical Data in RDF/Ontolex - All languages, version 20260601," Zenodo. https://doi.org/10.5281/zenodo.21496349 (concept record https://doi.org/10.5281/zenodo.10040982). Read October 26, 2026.
15. MorphyNet project, "MorphyNet: README.md," GitHub repository kbatsuren/MorphyNet. https://github.com/kbatsuren/MorphyNet. Read October 26, 2026.
16. National Library of Medicine, Lexical Systems Group, "The SPECIALIST Lexicon," lhncbc.nlm.nih.gov. https://lhncbc.nlm.nih.gov/LSG/Projects/lexicon/current/web/index.html. Read October 26, 2026.
17. National Library of Medicine, Lexical Systems Group, "SPECIALIST Lexicon: Word Count Stats," lhncbc.nlm.nih.gov. https://lhncbc.nlm.nih.gov/LSG/Projects/lexicon/current/docs/designDoc/UDF/wordStats/index.html. Read October 26, 2026.
18. National Library of Medicine, Lexical Systems Group, "Terms and Conditions for Use of the SPECIALIST NLP Tools," lhncbc.nlm.nih.gov. https://lhncbc.nlm.nih.gov/LSG/Projects/lexicon/current/web/termsAndConditions.html. Read October 26, 2026.
19. Carnegie Mellon University Speech Group, "The CMU Pronouncing Dictionary," speech.cs.cmu.edu. http://www.speech.cs.cmu.edu/cgi-bin/cmudict. Read October 26, 2026.
20. Carnegie Mellon University, "cmudict: README and LICENSE," GitHub repository cmusphinx/cmudict. https://github.com/cmusphinx/cmudict. Read October 26, 2026.
21. K. Atkinson, "SCOWLv1 Readme," wordlist.aspell.net, commit b22230c of July 22, 2024. https://wordlist.aspell.net/scowl_v1-readme/. Read October 26, 2026.
22. K. Atkinson, "News – ESDB," wordlist.aspell.net. https://wordlist.aspell.net/news/. Read October 26, 2026.
23. K. Atkinson and others, "Copyright," SCOWL release 2020.12.07, GitHub repository en-wl/wordlist, tag rel-2020.12.07. https://raw.githubusercontent.com/en-wl/wordlist/rel-2020.12.07/scowl/Copyright. Read October 26, 2026.
24. K. Atkinson, "English Speller Database (ESDB): README.md," GitHub repository en-wl/wordlist, branch v2. https://github.com/en-wl/wordlist/blob/v2/README.md. Read October 26, 2026.

25. K. Atkinson, "Word Sources and Tags (docs/sources.md)," GitHub repository en-wl/wordlist, branch v2. https://github.com/en-wl/wordlist/blob/v2/docs/sources.md. Read October 26, 2026.

26. K. Atkinson, "English Speller Database (ESDB)," home page. https://wordlist.aspell.net/. Read October 26, 2026.

27. K. Atkinson, "Copyright," GitHub repository en-wl/wordlist, branch v2. https://github.com/en-wl/wordlist/blob/v2/Copyright. Read October 26, 2026.

28. LibreOffice contributors, "README_en_US.txt, README_en_CA.txt, README_en_AU.txt, and README_en_GB.txt," GitHub repository LibreOffice/dictionaries, folder en, branch master. https://github.com/LibreOffice/dictionaries/tree/master/en. Read October 26, 2026.

29. M. Cooper and A. Beale, "ENABLE (README-enable2k.txt)," mirrored in GitHub repository BartMassey/wordlists, commit 2845b64. https://raw.githubusercontent.com/BartMassey/wordlists/2845b642405f6e63db0af966a8260e99366d5b3e/README-enable2k.txt. Read October 26, 2026.

30. A. Beale, "Version 6 of the 12dicts word lists (readme, release 6.0.2); 12Dicts Package," wordlist.aspell.net. https://wordlist.aspell.net/12dicts-readme/ and https://wordlist.aspell.net/12dicts/. Read October 26, 2026.

31. G. Ward, "Moby Word Lists (Moby Words II documentation notes)," Project Gutenberg eBook 3201, released May 1, 2002. https://www.gutenberg.org/ebooks/3201. Read October 26, 2026.

32. G. Ward, "Moby Part of Speech List (Moby Part-of-Speech II documentation notes)," Project Gutenberg eBook 3203. https://www.gutenberg.org/ebooks/3203. Read October 26, 2026.

33. M. L. Cooper; rehosted by A. B. Schaefer, "Yet Another Word List (YAWL): README.md," GitHub repository elasticdog/yawl. https://github.com/elasticdog/yawl. Read October 26, 2026.

34. lorenbrichter/Words maintainers, "Words: README.md and LICENSE," GitHub repository lorenbrichter/Words. https://github.com/lorenbrichter/Words. Read October 26, 2026.

35. Wordnik, "wordlist: README.md and LICENSE," GitHub repository wordnik/wordlist. https://github.com/wordnik/wordlist. Read October 26, 2026.

36. dwyl, "english-words: README.md, LICENSE.md, and repository description," GitHub repository dwyl/english-words. https://github.com/dwyl/english-words. Read October 26, 2026.

37. J. A. Woods and the FreeBSD Project, "README (share/dict)," FreeBSD source tree. https://cgit.freebsd.org/src/plain/share/dict/README. Read October 26, 2026.

38. Library of Congress, Copyright Office, "U.S. Copyright Renewals, 1961 July - December," Project Gutenberg eBook 11824, a transcription of the Catalog of Copyright Entries, lines 26239–26256. https://www.gutenberg.org/ebooks/11824. Read October 26, 2026.

39. R. Speer, "wordfreq: README.md," GitHub repository rspeer/wordfreq, branch master. https://github.com/rspeer/wordfreq. Read October 26, 2026.

40. R. Speer, "Why wordfreq will not be updated (SUNSET.md)," GitHub repository rspeer/wordfreq, branch master. https://github.com/rspeer/wordfreq/blob/master/SUNSET.md. Read October 26, 2026.

41. Python Package Index, "wordfreq," project metadata. https://pypi.org/project/wordfreq/. Read October 26, 2026.

42. Google Books Ngram Viewer team, "Google Books Ngram Exports," storage.googleapis.com. https://storage.googleapis.com/books/ngrams/books/datasetsv3.html. Read October 26, 2026.

43. J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, W. Brockman, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden, "Quantitative analysis of culture using millions of digitized books," Science, vol. 331, no. 6014, pp. 176–182, 2011. https://doi.org/10.1126/science.1199644.

44. Y. Lin, J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, and S. Petrov, "Syntactic Annotations for the Google Books Ngram Corpus," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, System Demonstrations, pp. 169–174, 2012. https://aclanthology.org/P12-3029/.

45. Google Books Ngram Viewer team, "Google Books Ngram Viewer: About," books.google.com. https://books.google.com/ngrams/info. Read October 26, 2026.

46. E. A. Pechenick, C. M. Danforth, and P. S. Dodds, "Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution," PLoS ONE, vol. 10, no. 10, e0137041, 2015. https://doi.org/10.1371/journal.pone.0137041.

47. M. Brysbaert and B. New, "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, vol. 41, no. 4, pp. 977–990, 2009. https://doi.org/10.3758/BRM.41.4.977. (Accepted manuscript read.)

48. Ghent University, Department of Experimental Psychology, "SUBTLEXus," ugent.be. https://www.ugent.be/pp/experimentele-psychologie/en/research/documents/subtlexus. Read October 26, 2026.

49. M. Brysbaert, "SUBTLEX_US word frequency database (with license.txt of May 7, 2024)," OSF project djpqz. https://osf.io/djpqz/. Read October 26, 2026.

50. W. J. B. van Heuven, P. Mandera, E. Keuleers, and M. Brysbaert, "SUBTLEX-UK: A new and improved word frequency database for British English," Quarterly Journal of Experimental Psychology, vol. 67, no. 6, pp. 1176–1190, 2014. https://doi.org/10.1080/17470218.2013.850521. (Author manuscript read.)

51. M. Brysbaert, "SUBTLEX-UK word frequencies (with license.txt of May 7, 2024)," OSF project zq49t. https://osf.io/zq49t/. Read October 26, 2026.

52. D. A. Balota, M. J. Yap, M. J. Cortese, K. A. Hutchison, B. Kessler, B. Loftis, J. H. Neely, D. L. Nelson, G. B. Simpson, and R. Treiman, "The English Lexicon Project," Behavior Research Methods, vol. 39, no. 3, pp. 445–459, 2007. https://doi.org/10.3758/BF03193014.

53. English Lexicon Project, Washington University in St. Louis, "Home page; About The English Lexicon Project," elexicon.wustl.edu. https://elexicon.wustl.edu/. Read October 26, 2026, through the Internet Archive captures of October 14, 2026 (home page) and August 25, 2019 (About page).

54. T. Brants and A. Franz, "Web 1T 5-gram Version 1 (LDC2006T13)," Linguistic Data Consortium catalog, 2006. https://catalog.ldc.upenn.edu/LDC2006T13. Read October 26, 2026.

55. T. Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for Datasets," arXiv:1803.09010v8, December 1, 2021; published in Communications of the ACM, December 2021. https://arxiv.org/abs/1803.09010.

56. E. M. Bender and B. Friedman, "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science," Transactions of the Association for Computational Linguistics, vol. 6, pp. 587–604, 2018. https://doi.org/10.1162/tacl_a_00041.

57. E. M. Bender, B. Friedman, and A. McMillan-Major, "A Guide for Writing Data Statements for Natural Language Processing" (schema version 2), University of Washington Tech Policy Lab, 2021. https://techpolicylab.uw.edu/wp-content/uploads/2021/11/Data_Statements_Guide_V2.pdf.

58. A. McMillan-Major and E. M. Bender, "Data Statements for Natural Language Processing: README.md and template-v3.md (schema version 3, 2024)," GitHub repository TechPolicyLab/Data-Statements. https://github.com/TechPolicyLab/Data-Statements. Read October 26, 2026.

59. M. Pushkarna, A. Zaldivar, and O. Kjartansson, "Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI," arXiv:2204.01075v1, 2022 (ACM FAccT 2022). https://arxiv.org/abs/2204.01075.

60. Hugging Face, "datasetcard_template.md," GitHub repository huggingface/huggingface_hub, branch main.

https://github.com/huggingface/huggingface_hub/blob/main/src/huggingface_hub/templates/datasetcard_template.md. Read October 26, 2026.

61. M. Akhtar, O. Benjelloun, C. Conforti, L. Foschini, and others, "Croissant: A Metadata Format for ML-Ready Datasets," in Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Datasets and Benchmarks Track, 2024; arXiv:2403.19546v3. https://arxiv.org/abs/2403.19546.

62. O. Benjelloun and E. Simperl, editors, "Croissant Format Specification, Version 1.1 (published January 29, 2026)," MLCommons, GitHub repository mlcommons/croissant. https://github.com/mlcommons/croissant/blob/main/docs/croissant-spec-1.1.md. Read October 26, 2026.

63. MLCommons, "Croissant Format Specification, Version 1.0 (published March 1, 2024)," GitHub repository mlcommons/croissant. https://github.com/mlcommons/croissant/blob/main/docs/croissant-spec.md. Read October 26, 2026.

64. M. Akhtar, N. Jain, J. Giner-Miguelez, O. Benjelloun, E. Simperl, L. Aroyo, R. Shinde, and M. Kuchnik, "Croissant RAI Specification, Version 1.0 (published March 6, 2024)," MLCommons, GitHub repository mlcommons/croissant. https://github.com/mlcommons/croissant/blob/main/docs/croissant-rai-spec.md. Read October 26, 2026.

65. NeurIPS 2026, "Call For Evaluations & Datasets 2026," neurips.cc. https://neurips.cc/Conferences/2026/CallForEvaluationsDatasets. Read October 26, 2026.

66. M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, and others, "The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data, vol. 3, 160018, 2016. https://doi.org/10.1038/sdata.2016.18.

67. TEI Lex-0 contributors, "TEI Lex-0: A baseline encoding for lexicographic data, version 0.9.5 (February 8, 2026)," lex-0.org, chapter 1, "Header". https://lex-0.org/ and https://lex-0.org/header.html. Read October 26, 2026.

68. T. Tasovac, L. Romary, E. Tóth-Czifra, and I. Marinski, "D6.4 Lexicographic Data Seal of Compliance," ELEXIS project deliverable, July 30, 2021. https://elex.is/wp-content/uploads/2021/08/ELEXIS_D6_4_Lexicographic_Data_Seal_of_Compliance.pdf.

69. P. Labropoulou, K. Gkirtzou, M. Gavriilidou, M. Deligiannis, and others, "Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid," in Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pp. 3428–3437, 2020. https://aclanthology.org/2020.lrec-1.420/.

70. N. Calzolari, R. Del Gratta, G. Francopoulo, J. Mariani, F. Rubino, I. Russo, and C. Soria, "The LRE Map. Harmonising Community Descriptions of Resources," in Proceedings of the Eighth

International Conference on Language Resources and Evaluation (LREC 2012), 2012. https://aclanthology.org/L12-1450/.

71. J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner, "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus," arXiv:2104.08758v2, 2021 (EMNLP 2021). https://arxiv.org/abs/2104.08758.

72. Y. Elazar, A. Bhagia, I. Magnusson, A. Ravichander, D. Schwenk, A. Suhr, P. Walsh, D. Groeneveld, L. Soldaini, S. Singh, H. Hajishirzi, N. A. Smith, and J. Dodge, "What's In My Big Data?," in Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024); arXiv:2310.20707v2. https://arxiv.org/abs/2310.20707.

73. A. Rauber, A. Asmi, D. van Uytvanck, and S. Proell, "Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC)," Research Data Alliance, October 20, 2015. https://doi.org/10.15497/RDA00016.

74. H. N. Andreassen, A. L. Berez-Kroeker, L. Collister, P. Conzett, C. Cox, K. De Smedt, B. McDonnell, and the Research Data Alliance Linguistic Data Interest Group, "Tromsø recommendations for citation of research data in linguistics," Research Data Alliance, December 4, 2019. https://doi.org/10.15497/RDA00040.

75. M. Gavrilidou, P. Labropoulou, E. Desipri, S. Piperidis, H. Papageorgiou, M. Monachini, F. Frontini, T. Declerck, G. Francopoulo, V. Arranz, and V. Mapelli, "The META-SHARE Metadata Schema for the Description of Language Resources," in Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), 2012. https://aclanthology.org/L12-1593/.

76. O. Sainz, J. A. Campos, I. García-Ferrero, J. Etxaniz, O. Lopez de Lacalle, and E. Agirre, "NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark," in Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 10776–10787, 2023. https://doi.org/10.18653/v1/2023.findings-emnlp.722.

77. C. Xu, S. Guan, D. Greene, and M-T. Kechadi, "Benchmark Data Contamination of Large Language Models: A Survey," arXiv:2406.04244v1, 2024. https://arxiv.org/abs/2406.04244.

78. C. Deng, Y. Zhao, Y. Heng, Y. Li, J. Cao, X. Tang, and A. Cohan, "Unveiling the Spectrum of Data Contamination in Language Models: A Survey from Detection to Remediation," in Findings of the Association for Computational Linguistics: ACL 2024, pp. 16078–16092, 2024. https://doi.org/10.18653/v1/2024.findings-acl.951.

79. W. Shi, A. Ajith, M. Xia, Y. Huang, D. Liu, T. Blevins, D. Chen, and L. Zettlemoyer, "Detecting Pretraining Data from Large Language Models," in Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024); arXiv:2310.16789v3. https://arxiv.org/abs/2310.16789.

80. S. Golchin and M. Surdeanu, "Time Travel in LLMs: Tracing Data Contamination in Large Language Models," in Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024); arXiv:2308.08493v3. https://arxiv.org/abs/2308.08493.

81. Y. Oren, N. Meister, N. Chatterji, F. Ladhak, and T. B. Hashimoto, "Proving Test Set Contamination in Black Box Language Models," arXiv:2310.17623v2, 2023. https://arxiv.org/abs/2310.17623.

82. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, and others, "Language Models are Few-Shot Learners," arXiv:2005.14165v4, 2020. https://arxiv.org/abs/2005.14165.

83. N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramèr, and C. Zhang, "Quantifying Memorization Across Neural Language Models," in Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023); arXiv:2202.07646v3. https://arxiv.org/abs/2202.07646.

84. A. Srivastava, A. Rastogi, A. Rao, and others, "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models," Transactions on Machine Learning Research, 2023; arXiv:2206.04615v3. https://arxiv.org/abs/2206.04615.

85. D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman, "GPQA: A Graduate-Level Google-Proof Q&A Benchmark," arXiv:2311.12022v1, 2023. https://arxiv.org/abs/2311.12022.

86. J. Wei, Z. Sun, S. Papay, S. McKinney, J. Han, I. Fulford, H. W. Chung, and others, "BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents," arXiv:2504.12516v1, 2025. https://arxiv.org/abs/2504.12516.

87. G. Leech, J. J. Vazquez, N. Kupper, M. Yagudin, and L. Aitchison, "Questionable practices in machine learning," arXiv:2407.12220v2, 2024. https://arxiv.org/abs/2407.12220.

88. Anthropic, "System Card: Claude Opus 4.5," November 2025, section 2.2, "Decontamination." https://www.anthropic.com/claude-opus-4-5-system-card.

89. A. Chowdhery, S. Narang, J. Devlin, and others, "PaLM: Scaling Language Modeling with Pathways," arXiv:2204.02311v5, 2022. https://arxiv.org/abs/2204.02311.

90. OpenAI, "GPT-4 Technical Report," arXiv:2303.08774v6, 2023 (revised March 4, 2024). https://arxiv.org/abs/2303.08774.

91. S. Chen, Y. Chen, Z. Li, Y. Jiang, Z. Wan, Y. He, D. Ran, T. Gu, H. Li, T. Xie, and B. Ray, "Recent Advances in Large Language Model Benchmarks against Data Contamination: From Static to Dynamic Evaluation," arXiv:2502.17521v2, 2025. https://arxiv.org/abs/2502.17521.

92. Z. Han, M. Mankikar, J. Michael, and Z. Wang, "Search-Time Data Contamination," arXiv:2508.13180v1, 2025. https://arxiv.org/abs/2508.13180.

93. T. Ishida, T. Lodkaew, and I. Yamane, "CapBencher: Give Your LLM Benchmark a Built-in Alarm for Test-Set Overfitting," in Proceedings of the 43rd International Conference on Machine Learning, 2026; arXiv:2505.18102v7. https://arxiv.org/abs/2505.18102.

94. N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks," arXiv:1802.08232v3, 2019. https://arxiv.org/abs/1802.08232.

95. A. Jacovi, A. Caciularu, O. Goldman, and Y. Goldberg, "Stop Uploading Test Data in Plain Text: Practical Strategies for Mitigating Data Contamination by Evaluation Benchmarks," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 5075–5084, 2023. https://doi.org/10.18653/v1/2023.emnlp-main.308.

96. O. Sainz, I. García-Ferrero, A. Jacovi, J. A. Campos, Y. Elazar, E. Agirre, Y. Goldberg, and others, "Data Contamination Report from the 2024 CONDA Shared Task," arXiv:2407.21530v2, 2024. https://arxiv.org/abs/2407.21530.

97. N. Jain, K. Han, A. Gu, W.-D. Li, F. Yan, T. Zhang, S. Wang, A. Solar-Lezama, K. Sen, and I. Stoica, "LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code," arXiv:2403.07974v2, 2024. https://arxiv.org/abs/2403.07974.

98. C. White, S. Dooley, M. Roberts, A. Pal, and others, "LiveBench: A Challenging, Contamination-Limited LLM Benchmark," in Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025); arXiv:2406.19314v2. https://arxiv.org/abs/2406.19314.

99. H. Deng, W. Jiao, X. Liu, M. Zhang, and Z. Tu, "NewTerm: Benchmarking Real-Time New Terms for Large Language Models with Annual Updates," in Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Datasets and Benchmarks Track; arXiv:2410.20814v1. https://arxiv.org/abs/2410.20814.

100. ryanjosephkamp, "The CEOL Universe, version 1.0," dataset card, Hugging Face dataset ryanjosephkamp/ceol-universe, built October 25, 2026; the card's source is `pipeline/research/release/CARD_v1.0.md` in https://github.com/ryanjosephkamp/ceol-universe. Read October 26, 2026.

101. H. Jung and H. Gonen, "PhantomBench: Benchmarking the Non-existential Threat of Language Models," arXiv:2606.11105v1, 2026. https://arxiv.org/abs/2606.11105.

102. C. Uluoglakci and T. Taskaya Temizel, "HypoTermQA: Hypothetical Terms Dataset for Benchmarking Hallucination Tendency of LLMs," in Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pp. 95–136, 2024. https://doi.org/10.18653/v1/2024.eacl-srw.9.

103. G. Martínez, J. Conde, E. Merino-Gómez, B. Bermúdez-Margaretto, J. A. Hernández, P. Reviriego, and M. Brysbaert, "Establishing Vocabulary Tests as a Benchmark for Evaluating

Large Language Models," arXiv:2310.14703v2, 2024. https://arxiv.org/abs/2310.14703.