AccStack: An agentically compiled catalog of word stress and tone, with sources

Matthew D. Schwartz* Kevin M. Ryan† Matthew K. Gordon‡ Joe Pater§
Larry M. Hyman¶

October 1, 2026

This work is preliminary. Some content remains AI-generated and is still being validated.

PRELIMINARY

Abstract

This paper presents AccStack, a catalog of word stress and tone at https://accstack.org, which records what published descriptions say about word stress and tone for 6,070 languages. Each entry gives the stress type, the stressed syllables at each word length the source states, the exceptions and domain of the rule, and the tone system where the source describes one. The catalog was compiled by a large language model, which read the sources under our written instructions, quoted the passage that decides each entry and coded it. A catalog of this size is beyond what a linguist or a small group could compile by hand in any reasonable time, and therefore not every entry has been checked by hand. However, every entry was made to the same standard, and each shows its citation, the quoted passage and the reason for the ruling, so that readers can judge the interpretation for themselves and corrections are easy to make.

1 Introduction

AccStack records the statements that descriptions of languages make about the position of stress in a word and about whether they have tone. It has an entry for 6,070 languages. Each entry gives a coding of the language's stress and tone together with the passage of the description that the coding rests on, so that the one can be checked against the other. The entries were compiled by a large language model (LLM) working under our written instructions. A companion bibliography lists works on each language and what each says about stress or tone. The catalog and the bibliography are on https://accstack.org, where every entry is shown with its fields and both can be downloaded.

Compilations of such statements are the material of the typology of word stress and tone. Hyman (1977) tallied the position of main stress in 444 descriptions, and Gordon (2002) surveyed

*Department of Physics, Harvard University, Cambridge, MA 02138, USA. Corresponding author: schwartz@physics.harvard.edu.

†Department of Linguistics, Harvard University, Cambridge, MA 02138, USA.

‡Department of Linguistics, University of California, Santa Barbara, CA 93106, USA.

§Department of Linguistics, University of Massachusetts Amherst, Amherst, MA 01003, USA. Work supported by NSF grant BCS-2140826 to the University of Massachusetts Amherst; results and views are not endorsed by the NSF

¶Department of Linguistics, University of California, Berkeley, CA 94720, USA.

262 languages with fixed stress. StressTyp (Goedemans & van der Hulst 2009) provided the codings that the stress chapters of the *World Atlas of Language Structures* (WALS) present (Goedemans & van der Hulst 2013a,b,c,d). StressTyp was merged with the Stress Pattern Database (Heinz 2009) into StressTyp2 (Goedemans et al. 2014), which covers 699 languages in its archived version. LAPSyD (Maddieson et al. 2013), chapter 13 of WALS (Maddieson 2013), PHOIBLE (Moran & McCloy 2019), Maddieson & Benedict (2024) and ThoTDB (Maslinsky et al. 2025) record whether a language has tone. Two of them record a stress category as well. The catalog builds on this work: it reuses the codings and references of StressTyp2 and of Gordon's survey, and LLM readings of the literature made for this catalog supply the rest of its entries.

An entry gives the language's stress type from a closed list, the primary and secondary stresses at each word length where the source states them, the exceptions and the domain of the rule. It also records the variety described, the tone system where the source describes one, and a confidence level. The deciding passage is quoted verbatim from the source with the page or section and a translation where needed. Where the reading of a page differs from an earlier database's coding, or two descriptions of a language disagree, both are shown. Section 4.1 lists the fields.

Section 2 defines the terms used and describes the earlier surveys and databases. Sections 3 to 7 describe how the catalog was compiled, what it and the bibliography contain, what checks were made on the entries, how far its codings agree with the earlier databases and where its limits lie.

PRELIMINARY

2 Earlier surveys and databases

The earlier cross-linguistic surveys and databases of word stress were all compiled from the descriptive literature. For Gordon's survey and StressTyp2, whose fixed-stress entries the catalog incorporates, we note whether an entry gives page, quotation, variety, secondary stress and domain.

2.1 Word stress, pitch accent and tone

By word stress we mean a prominence that a description assigns to one syllable of each word as a property of the word rather than of the phrase, obligatory and culminative in the sense of Hyman (2006). Secondary stresses may fall on other syllables, though many grammars say nothing about them. A grammar will often state the rule in a sentence or two, naming the syllable that carries the main stress and saying what, if anything, may move it. In the usual terms (Hayes 1995; van der Hulst 2014), stress is *fixed*, or quantity-insensitive, when its position follows from the word's edges and number of syllables alone. Three fixed-stress rules are set out in (1), where $\acute{\sigma}$ marks primary stress, $\grave{\sigma}$ secondary stress and σ an unstressed syllable, following the acute and grave accents of Gordon (2002: 493).

(1) Three fixed-stress rules, at word lengths two to five

a. initial stress, as in Hungarian [hung1274] (Uralic) and Chechen [chec1245] (Nakh-Daghestanian): $\acute{\sigma}\sigma$, $\acute{\sigma}\sigma\sigma$, $\acute{\sigma}\sigma\sigma\sigma$, $\acute{\sigma}\sigma\sigma\sigma\sigma$

b. penultimate stress, as in Swahili [swah1253] (Atlantic-Congo) and Amara [amar1272] (Austronesian): $\acute{\sigma}\sigma$, $\sigma\acute{\sigma}\sigma$, $\sigma\sigma\acute{\sigma}\sigma$, $\sigma\sigma\sigma\acute{\sigma}\sigma$

c. initial main stress with a secondary stress on every second syllable after it, never on the last, as in Badimaya [badi1246] (Pama-Nyungan) and Karelian [kare1335] (Uralic): $\acute{\sigma}\sigma$,

σσ́σ, *σσ̀σ*, *σσ̀σσ*

Stress is *weight-sensitive* when a heavy syllable, such as one with a long vowel or a closing consonant, can attract it away from the position a fixed-stress rule would give it. A weight-sensitive system is *unbounded* when stress can fall on a heavy syllable at any distance from the edge. Stress is *lexical* when the position differs across words of the same syllabic and morphological shape, and *morphological* when it is fixed relative to a root, stem or affix rather than the word. Descriptions of the same language sometimes classify them differently in these terms; for example, whether a system has lexical stress or predictable stress with exceptions can be a matter of judgment. Many descriptions report lexical tone or a pitch accent and say nothing about stress. Such a description may leave open whether the language has word stress at all (see de Lacy 2014).

In the catalog, a language described with both tone and a rule of stress is classed by the rule, with the tonal statement quoted beside it.

2.2 The surveys

Hyman (1977) tallied the dominant position of main stress in the 444 descriptions then available to him, weight-sensitive and alternating systems included (Hyman 1977: 41, 58; Gordon 2002: 494–495, footnote 2, Table II). Gordon's own survey covers 262 languages "with predictable quantity-insensitive stress patterns (i.e., predominantly non-phonemic with a clearly dominant pattern)" (Gordon 2002: 493). Each row of the survey's appendix gives the language, its class (Initial, Penult, "Odd-numbered from L minus final") and its family (2b below). The references were made available on Gordon's website (Gordon 2002: 494), and the appendix has no column for page, quotation, variety or domain.

2.3 StressTyp, the Stress Pattern Database and StressTyp2

StressTyp (Goedemans & van der Hulst 2009) coded 510 languages from grammars and typological studies. Its print companion is the survey edited by van der Hulst et al. (2010). Heinz's Stress Pattern Database (2007, 2009) combined Gordon's survey with an earlier compilation by Bailey (1995), coding each pattern's primary and secondary stresses explicitly. Goedemans, Heinz and van der Hulst then merged StressTyp and the Stress Pattern Database into StressTyp2 (2014). Its archived first version covers 699 languages, and the online version reports over 750. The compilers' stated aim is "to provide a key to the literature, and not to impose our own views on stress/accent" (Goedemans et al. 2014, documentation, §1).

A StressTyp2 entry links a language and its ISO 639-3 code to one or more lects (varieties), and each lect to its patterns, classed as primary, secondary, exceptional or one of several further kinds. A pattern carries a prose description, letter codes for the main and secondary stresses, a finite-state representation, a quantity-sensitivity attribute, example words and sources. Some entries also hold passages quoted from the sources, with page numbers. Example (2) shows one language's entry in StressTyp2 and its row in Gordon's appendix.

(2) Pintupi-Luritja [pint1250] (Pama-Nyungan) in StressTyp2 and in the survey of Gordon (2002)

a. StressTyp2: language *Pintupi-Luritja*, ISO 639-3 `piu`; primary stress code `I`; secondary stress code `i2@1L, no 1R`

b. Gordon (2002), appendix: language *Pintupi*; class "Odd-numbered from L minus final"; family "Australian"

PRELIMINARY

Both code main stress on the first syllable and secondary stress on every odd-numbered syllable after it except the last: *σσσσσ*, *σσσσσσ*. The StressTyp2 primary code names the main-stress syllable (`I` initial, `S` second, `T` third, `A` antepenultimate, `P` penultimate, `U` final). The secondary code places a stress on every second syllable from the first (`i2@1L`) and none on the final (`no 1R`). A StressTyp2 entry thus names the variety and codes secondary stress, and gives page and quotation where a passage has been entered. A field for the domain of a pattern (root, stem or phrase) exists but appears to be filled in for few entries in the archived version. Gordon's survey has no such field.

2.4 WALS, LAPSyD and the databases with a tone field

Four chapters of the *World Atlas of Language Structures* (WALS) present StressTyp's data as one value per language per map (Goedemans & van der Hulst 2013a,b,c,d). They map fixed stress locations (chapter 14), weight-sensitive stress (15), weight factors (16) and rhythm types (17), each value coded from the fuller StressTyp entry. LAPSyD (Maddieson et al. 2013), which covers 764 languages, stems from the UPSID and WALS databases and adds syllable structure, stress and tone. Each language is assigned a stress category (none, minimal or lexical) and a tone category (none, marginal, simple, moderately complex or complex), with a note on each and references for the items relied on.

The databases below, like LAPSyD, have a tone field, and one of them has a stress field as well. Chapter 13 of WALS (Maddieson 2013) classes 527 languages as having no tones, a simple tone system or a complex one. PHOIBLE 2.0 (Moran & McCloy 2019) holds 3,020 phoneme inventories for 2,186 languages, tones included where the inventory's source describes them. Maddieson & Benedict (2024) give a five-way tone category and a three-way stress category for 1,003 languages. ThoTDB (Maslinsky et al. 2025) marks each Glottolog language as tonal, toneless or unknown. For a planned sample of up to 250 languages it records the tone-bearing unit, the tonal levels, the tonemes with the criteria for each, downstep and the tonal rules. The catalog's own tone fields are part of the coding scheme described below, and the catalog's agreement with each of these databases is measured in Section 6.

3 Compilation of the catalog

The catalog was built from the Glottolog list of spoken languages and a list of candidate sources for each language. A large language model opened an accessible source from a language's list, quoted what it says about stress or tone and coded it. The quotation and the coding were then checked against the source in a second LLM reading.

3.1 Languages and candidate sources

The languages surveyed are the spoken languages that Glottolog lists at language level (Hammarström et al. 2025, version 5.2). From Glottolog's language-level list we leave aside the sign languages, pidgins, argots and ritual languages, and artificial languages, together with the provisional and unattested languoids that Glottolog keeps for bookkeeping. The remaining 7,803 languages, unclassified and mixed languages included, form the list against which coverage is measured (the frame). Entries are keyed to Glottolog's language codes, one entry per language, and a dialect or variety described under its own name is entered under its language. A few entries fall outside the frame but are kept: described pidgins, speech registers, one language Glottolog classes

PRELIMINARY

as unattested, and artificial languages such as Esperanto and Klingon, which the site lists only when the reader asks for them. Sign languages are not entered, since what their descriptions call stress is manual prominence.

A language's candidate sources begin with every grammar, grammar sketch and phonology in Glottolog's bibliography, and every other work there whose title names phonology, phonetics, sounds, syllables, tone, pitch, prosody, stress or accent. Version 5.1 of that bibliography lists 54,586 such works on 5,468 languages. To these are added the works cited in StressTyp2 (Goedemans et al. 2014), in the survey of Gordon (2002) and in the WALS stress chapters (Goedemans & van der Hulst 2013a,d), and the works found by an LLM while reading.

Each work is also given an access class. The class is the most accessible one supported by a copy on file, a library look-up (Harvard's catalog, HathiTrust, the Internet Archive), a wider copy search (Section 4.6), the reading itself or an open link recorded by Glottolog. A work is *digital, open* if an open copy was read or is on file, or an open link to one is recorded. It is *digital, not open* if a digital copy exists only behind a sign-in, a paywall, a search-only or lending view, a restricted repository or a dead link. Such a copy is cataloged but was not read. *Scan and deliver* means that no digital copy is known and a research library holds a print copy from which pages can be scanned on request. *Print only* means that the work is located, as a library or publisher's record, with no digital copy and no such holding, and *unknown* that it was not located or not yet checked. The scan-and-deliver and print-only classes reflect one library's holdings (Harvard's) at the time of the release and would probably differ elsewhere.

3.2 The reading procedure

The readings were made from August to early October 2026 by large language models of Anthropic's Claude family, run as agents, that is, with tools for fetching, saving and searching documents and left to decide for themselves, within our instructions, which works to open and which pages to read. Works that could not be downloaded automatically were fetched by hand through our library sign-ins. We wrote the coding rules, the classification of exceptions and the three cuts of Section 4.3, the access classes, the wording of the confidence scale and the reconciliation procedure, including how each kind of disagreement is settled. All the readings, reviews and confidence judgments described in this section and in Section 5 were produced by model readers applying those rules, under the instructions excerpted in Appendix A. We have spot-checked entries for languages we know, but we have not checked every entry and cannot attest that every one is correct (Section 7). Every change of classification and every data release is our decision. Most of the sources are open access and freely available online; some were read from university library collections. The entries quote short attributed passages to document a classification, and no source is redistributed.

Each entry compiled for the catalog rests on two separate readings by a language model, as does each statement about stress that the bibliography lists for a work. The second reading was made with the first report in hand, under instruction to refute it. Where a copy of the source could be saved, as for most entries, both readings used the same copy. In the first reading the model located an accessible source, generally trying copies on file before the open web. The passage on stress was then found by searching for the terms for stress, accent, prominence and tone in the language of the work (e.g. *akzent*, *acento*). A work marked silent may therefore still treat prominence under another heading. The deciding passage was quoted verbatim from the original, with a translation where needed and the page or section. Each passage was given a reading from the closed list, with the rule restated and the source's statements on secondary stress, exceptions and domain noted.

PRELIMINARY

In the second reading the work was opened again at the cited page, where possible, to check that the passage appears there verbatim. The surrounding section was read, the first reading confirmed or corrected, and a check made that the work describes this language rather than a neighbor or namesake. Where the second reading did not simply confirm the first, the source was read a third time and the case decided field by field, with reasons; both earlier readings were kept.

Once a language had an entry, the other open works listed for it were read in the same way and their relation to the entry recorded. A work *confirms*, *refines* (adding detail without changing the placement or type of stress), *contradicts* (the difference being stated), is *silent*, is *not about this language*, or was *not opened*, with the reason. Example (3) shows one entry.

(3) Seimat [seim1238] (Austronesian; Wozna & Wilson 2005: 7)

Source	*Seimat Grammar Essentials*, §1.3 'Stress', p. 7
Quotation	"There is no evidence of contrastive stress in Seimat. Stress normally occurs on the penultimate syllable of a word, except in a small group of trisyllabic words where reduplication occurs. In this case the primary stress occurs on the antepenult and secondary stress on the ultima. Examples: /ˈpaŋaˌpaŋ/ 'moon' /ˈsuhuˌsuh/ 'Triton trumpet shell' "
Reading	fixed penultimate, *σ́σ σσ́σ σσσ́σ*; secondary stress stated only for the reduplicated trisyllables
Domain	the word; exceptions a listed set (class: compounds or reduplications)
Type, grade	fixed stress, under the core-vocabulary cut; open text, first-hand
Check	quotation found verbatim at the cited page; reading and language identity confirmed
Confidence	fairly confident: one first-hand description, checked against the source

3.3 The coding scheme

Each work consulted is assigned one reading from the closed list of Appendix A.2 (a stress position or type of system, or no usable statement), and the language one of the types of Table 2. For fixed stress, the rule field gives the stressed syllables at each word length from two to eight syllables where the source states a general rule. Otherwise it gives them only at the lengths the source states or exemplifies, without marking which lengths rest on cited forms. Rules are written here in the notation of example (1), *σ́σ*, *σσ́σ*, …, and in the released data with one digit per syllable (`1` stressed), so that `10 010 0010` is penultimate stress in words of two to four syllables. Secondary stress has its own field, with its quotation, source and status: a rule given, present but not placed, one stress per word, found only in compounds, reduplications or phrases, not mentioned, no word stress, or source not seen. A secondary stress that the source gives as optional, variable or morphologically placed, or states for some word lengths only, is described in words and not marked in the rule. For fixed stress the classes of exception that the source states are coded, with the forms it cites. The other fields of Table 1, each with the source's own sentence, cover the weight criterion, the domain (word, phrase, word and phrase under one rule, root, or unstated), affixes and clitics, vowel length and tone. Where the source describes tone or pitch accent, or states that the language has neither, the entry records the type of tone system, the number of tones if stated, and the passage and page. The tone statement is filled in 2,845 entries. Three fields of the tone statement are filled only from the passage's own words. The relation between tone and stress is entered where the source states it, downstep where the source uses the word, and the tone-bearing unit where the source names it.

PRELIMINARY

A further field gives the kind of evidence the primary source offers, and whether its data are its own or restated from another work. The entry also names the variety described and its relation to the Glottolog language, keeping the code only if the second reading confirmed it, and says whether any StressTyp2 or WALS coding matches our reading of the source.

Some statements are entered without being coded as word stress. A description that reports tone or pitch accent and says nothing of word stress is entered with that report alone (1,910 entries, Table 2). No stress position is inferred from it; whether the language also has word stress is left to what its descriptions state (see Hyman 2006). A statement of phrase-level prominence alone is entered as phrase-level only rather than as word stress (46 entries). An entry whose deciding passage is quoted from a database or other secondary work rather than the primary source is marked *carried quote* until the passage is found there (105 entries). Secondary stresses that a source does not mention are left unstated. Positions are coded by word length as the source states them, not analyzed into feet, so as not to depend on a metrical analysis the source may not share. A rule stated over moras is coded by the syllable that carries the main stress in words of light syllables, with a note that the source counts moras. Where the source also states that syllable weight moves the main stress, as for Ho-Chunk for example, the language is entered as weight-sensitive.

3.4 Evidence grades, confidence levels and conflicting descriptions

Deciding quotations are given one of five evidence grades, by the form in which the passage was read. Three can be first-hand: a *page image* (2,296 entries), the *publisher text* (635) and an *open text*, a repository or author's copy (2,789). The other two are a *snippet* from a search engine, with the work unopened (244), and a *carried quote*, taken second-hand from StressTyp2, WALS or a later work quoting the description (105). A quotation counts as first-hand only when it was read in a saved copy and the second reading found it there at the stated page. Whether a work states the pattern from its own data or restates another work is recorded separately from the grade.

Every entry is rated on a five-level confidence scale. *Very confident* is "clear consensus" and *fairly confident* is "low number of sources or mild disagreement but still a clear winner". *Moderately confident* is "some real disagreement but you think you know what the answer is", *disputed* is "an active debate with authoritative sources disagreeing" and *not confident* is "basically just guessing". Very confident, fairly confident and not confident are set by rule from the works the entry lists, and moderately confident and disputed only by a judged reading of the sources against the five definitions (Appendix A.3). The rule counts works, not citations: editions, reprints and variant citations of a work are merged. Works by the same first author count as one, as do a work and a restatement of it, and works on another variety are not counted. *Very confident* by rule requires at least two independent works, read first-hand and not restatements of another work, that agree on the type of system and state the same stress rule. It further requires no confirmed contradiction, no database coding that agrees only in part, and no known but unread description of a fuller kind than any read for the entry. *Fairly confident* by rule means that a statement of the pattern was confirmed or, in 166 entries, read only in a snippet that shows it, but that the conditions for very confident are not all met. Most often no second independent work states the same rule; in 230 entries the quoted source itself restates another work, and in others a fuller description is known but unread. *Not confident* by rule means that no statement of the recorded pattern, other than a database coding, was confirmed in a saved copy or shown by a snippet. An entry that the rule would otherwise rate very or fairly confident is rated by the judged reading instead when it has a confirmed contradiction or a database coding that agrees only in part, or when its own type

PRELIMINARY

is *descriptions disagree*, *not determinable* or *variable*. It is also referred to that reading when a counted work describes the language's stress as not determinable or variable, or proves not to describe the language.

When descriptions of a language disagree, we do not code a compromise. If none can be preferred, the language's type is *descriptions disagree*. Otherwise the preferred classification stands, with the dissenting source, page and words beside it as a contradiction. A contradicting statement read later is attached in the same way and leaves the coding unchanged until we decide the case. A changed classification is entered with the deciding source and page quoted, and the superseded reading stays visible in the entry.

4 The catalog and the bibliography

An entry in the catalog is built around a passage quoted from a description of the language. Its other fields classify the stress pattern and record what the source says about exceptions, secondary stress, the stress domain and tone. A bibliography released with the catalog lists the works known on each language. The catalog is released as one data file and the bibliography as an index with a row per language, a file for each language and a summary of the counts.

4.1 The fields of an entry

Table 1 lists the fields of an entry, which is identified by Glottolog code (Hammarström et al. 2025), and records its provenance: a database (Goedemans et al. 2014; Gordon 2002) or our own reading of the sources. The deciding quotation is given verbatim with work, page and the kind of copy read.[1]

The entry for Yimas (Lower Sepik; Foley 1991: 75–77), one of those from StressTyp2, illustrates the format. Among its sources it quotes the grammar the database cites, "Yimas stress is essentially predictable by a simple rule: assign primary stress to the first syllable of the word, and secondary stress to the third syllable, if the word is longer than three syllables. Stress is not phonemic" (Foley 1991: 75), with the forms in (4) from the database's own word list.

(4) Yimas [yima1243] (Lower Sepik; Foley 1991: 75–77; glosses as given in StressTyp2's word list)
   a. [ˈawtmayŋi] 'sugarcane', [ˈyamparan] 'stand up' *σ́σσ*
   b. [ˈwuraˌtakay] 'turtle', [ˈwankaˌnawi] 'insect (sp.)' *σ́σσ̀σ*
   c. [tɨˈmarmaŋ] 'possum' *σσ́σ* (epenthetic initial vowel, the grammar's stated exception)

The rule field gives primary stress on the first syllable and a secondary stress on the third in words of four or more syllables (4b), written as the digit string shown in Table 1. StressTyp2's own coding, alternation on odd-numbered syllables, is entered beside it. The domain is coded as the word on the evidence of the sentence quoted. Yimas falls under the dominant-pattern cut and no narrower one. Besides the exception in (4c), the grammar states secondary stresses on certain suffixes and free variation between the first two syllables in a few words.

[1] The quotation is empty where the source gives no stress passage. One entry from an earlier release, for Bangala, has none and is classed as not determinable. Of the 299 entries made from descriptions of tone, 290 quote only the tone statement.

PRELIMINARY

Table 1: The fields of a catalog entry, with the number of entries in which each is filled where it applies.

Field	Contents	Entries filled
Language, provenance	Glottolog code, name, family, macroarea, coordinates; origin of the entry; relation of the variety described to the Glottolog language (the same 3,609, a dialect or variety 1,319, a wider group 133, another languoid 70)	6,070
Quotation, source	deciding passage verbatim, with work and page or section; translation where needed	5,779; 348
Evidence grade	page image, open text, publisher text, snippet, carried quote (Section 3.4)	6,069
Evidence in the source	described elicitation 2,516, impressionistic statement 2,305, phonological alternations 1,566, restatement 773, instrumental measurement 562, corpus or dictionary count 323, perception experiment 34 (one or more); own data 4,843, cites another work 363, unclear 458	5,664
Stress type	one of the twelve values of Table 2	6,070
Rule; rhythm class	a stress string for each word length, two to eight syllables, that the source states (`1` stressed, as in Yimas `10 100 1010 10100 101000 1010000 10100000`); single, two fixed, binary or ternary alternating, multiple, or as described	1,654 (1,609 fixed); 1,243
Secondary stress	with its own quotation: rule given 1,072, present but not placed 245, one stress per word 288, only in compounds, reduplications or phrases 211, not mentioned 3,058, no word stress 759, source not seen 48; later revision of the primary rule (203)	5,681
Domain	word 3,929, phrase 94, word and phrase 99, root 7, unstated 1,642; the source's sentence	5,771; 5,705
Exceptions	none, listed or pervasive; classes for the 1,610 fixed-stress entries (Section 4.3: none named 609, qualifying adverb only 131, loanwords 201, compounds 133, phrase level 157, secondary stress only 108, native lexical 477, morphological 135, clitics 57, other 310; one or more per entry); forms cited	5,719; 1,601; 2,022
Affixes, clitics	affixes counted by the rule, stress-bearing or outside its domain; clitics	2,754; 1,347
Weight, length, tone	weight criterion (CVV, CVV and CVC, CVC, full vs. reduced vowel, other); word edge; vowel-length contrast; tone or pitch-accent class; tone statement (Section 3.3)	994; 1,957; 4,451; 6,013; 2,845
Cuts	dominant-pattern, core-vocabulary and absolute cuts met	1,610; 953; 602
Confidence	level 4 (very confident) to 0 (not confident), with the sentence that states its reason	6,070
Contradictions; further sources	each description or check that disagrees, with its passage; cautions on thin evidence; database-coding notes; passages from other works read	427; 280; 333; 919
Database agreement	whether the StressTyp2 coding agrees (yes, partly, no); whether the WALS coding agrees (yes, partly, no)	330; 289
Samples	the WALS 100- and 200-language samples, a one-per-genus sample, and a sample maximally dispersed by genealogical and geographic distance	102; 202; 702; 996
Other fields	notes from later checks and on the variety (391); rule tested on the source's stress-marked words (200); audio clips (243) or links to recordings (4,175); works examined (median 7); Glottolog and source links	

Notes: counts separated by semicolons follow the order of the values in the row; a count in parentheses in the contents column is that of the value it follows and is not repeated in the last column.

PRELIMINARY

4.2 Languages and stress types

The release of 4 October 2026 holds 6,070 entries, one for each language, each identified by its Glottolog code. Of these languages, 6,008 belong to the frame of Section 3.1.[2] All but the 394 entries from StressTyp2 and Gordon's survey come from readings of the descriptive literature made for this catalog, beginning with a survey of open documentation (our first survey). The database entries and the sources they cite are compared in Section 5.2. Works found by searching the literature for descriptions of tone are the sources of 299 entries (261 of these record tone with no word-stress statement).

By stress type (Table 2) the largest class is tone or pitch accent with no word-stress statement, followed by fixed, quantity-insensitive stress. More than half of the entries carry a word-stress type, and under an eighth are assigned no type.

Table 2: Entries by stress type; subclasses indented.

Stress type	Entries
Tone or pitch accent; no word-stress statement	1,910
Fixed, quantity-insensitive stress	1,610
Weight-sensitive stress	750
stated as a rule	345
stated with hedges	405
Lexical stress	518
Morphologically assigned stress	421
No word stress	128
No type assigned	733
not determinable from the sources	556
descriptions disagree	50
segmentally conditioned	47
phrase-level only	46
variable	34
All entries	6,070

4.3 Fixed stress and stated exceptions

The sources of most fixed-stress entries qualify the rule in some way, and users of the catalog may reasonably differ on what should count as an exception. For the fixed-stress entries we therefore code the classes of exception their sources state, nine in all. Five, which we call peripheral, leave the primary stress of native simple words in citation form untouched. They are a qualifying adverb such as "usually" with no exception named, loanwords or names, compounds or reduplications, conditioning by the phrase or utterance, and exceptions in secondary stress only. The other four are native lexical items, morphological classes or affixes, clitics or particles, and other conditions such as weight in part of the vocabulary, free variation or dialect. The forms cited are entered with the class. An entry may carry several classes. Native lexical exceptions are the class most often stated, for about three in ten of the fixed-stress entries, and for more than a third the sources name none (Table 1).

[2]The other 62 entries are for artificial languages, pidgins, speech registers such as argots, and an unattested language.

PRELIMINARY

The catalog leaves the choice of admissible exceptions to the user and provides three nested cuts (Table 3). The *dominant-pattern* cut comprises every entry whose source states a quantity-insensitive positional rule as at least the predominant pattern, whatever exceptions it names. This is in effect Gordon's criterion, "predominantly non-phonemic with a clearly dominant pattern" (Gordon 2002: 493), and all fixed-stress entries meet it. The *core-vocabulary* cut keeps those whose rule is exceptionless for the core vocabulary, the stated exceptions being confined to the five peripheral classes. The *absolute* cut admits only entries whose rule is exceptionless without qualification of any kind. An entry whose source states no quantity-insensitive positional rule is entered not as fixed stress but under the type the source supports (morphological, weight-sensitive or lexical).

Table 3: The fixed-stress entries, all under the dominant-pattern cut, and the two exceptionless cuts, each contained in the row above.

	Entries	Rule strings	Complete rules	Distinct patterns
All fixed-stress entries	1,610	116	1,315	45
core-vocabulary cut	953	82	734	29
absolute cut	602	60	442	25

Notes: rule strings: distinct strings among the entries' rules; complete rules: entries whose rule covers every word length from three to eight syllables; distinct patterns: distinct complete rules at those lengths.

The secondary-stress and rule fields hold only what the source states. For 876 of the 1,610 fixed-stress entries, the source says nothing of secondary stress. It gives a rule for 463, states one stress per word for 104, confines secondary stress to compounds, reduplications or phrases for 87, and mentions it without placing it for 64. For the other 16 the field is blank or not determinable. Of the 1,609 entries with a fixed-stress rule, 294 state it for some word lengths only. Where a source states a general rule, the rule field gives it at every word length from two to eight syllables, so that at the longer lengths the string is the rule as stated rather than a cited form.

4.4 Confidence levels and recorded contradictions

Of the 6,070 entries, 988 are rated very confident, 4,051 fairly confident, 352 moderately confident, 84 disputed and 595 not confident. The confidence rule (Section 3.4) sets 4,099 of these ratings, and a judged reading of the sources sets the other 1,971; the two levels set only by judgment, moderately confident and disputed, together account for 436 entries. For most of the entries rated not confident, no usable statement of the recorded pattern was confirmed in a saved copy. In these entries the quotation is carried from another work, the source could not be reached, no statement or only a passing one was found, or the reading was unclear. The classifications of all entries rated not confident should be treated as provisional. Every entry gives the counts of works behind its rating, and on the site each count opens the list of those works. For 812 entries a description of a fuller kind than any read for the entry is known but not yet read; each such entry is marked *best evidence unread* and names the works wanted. For these and 193 other entries the site also lists 2,041 of the wanted works, with each work's access class and the pages or chapter to request. Of the 812 entries so marked, 96 meet every other condition for very confident. Four entries not so marked meet every condition except that the quoted work was read only in part. The confidence rule rates these 100 entries fairly confident rather than very confident, and a judged reading of the

PRELIMINARY

sources sets 15 of them at other levels.

When another description, the source's own examples or a later reading disagrees with an entry, the dissenting passage is attached with its citation; notes on database codings are kept apart and not counted as dissents. Table 1 gives the numbers of entries with a contradiction, with a caution on thin evidence and with a database-coding note; ten carry both a contradiction and a caution.

4.5 Coverage, samples and the languages without an entry

Coverage ranges from 69% in Papunesia to 89% in North America (Table 4, Figure 1). With each isolate and each language without a genealogical classification counted as a family of one, the languages with an entry belong to 461 families, 137 of them isolates. The families with the most entries are Atlantic-Congo (1,061), Austronesian (952), Indo-European (476), Sino-Tibetan (427), Afro-Asiatic (305) and Pama-Nyungan (197). Entries are also marked for membership in four balanced samples. Two are the 100- and 200-language samples of WALS (99 and 198 of the sample languages have an entry, a few of them under more than one Glottolog code). The others are a sample of one catalog language per WALS genus (702) and a maximally dispersed sample of 996 languages ranked by genealogical and geographic distance.

Table 4: Languages of the frame with an entry, by macroarea.

Macroarea	Frame languages	With an entry	Share
Papunesia	2,128	1,465	69%
South America	627	474	76%
Australia	358	275	77%
Africa	2,225	1,722	77%
Eurasia	1,773	1,459	82%
North America	692	613	89%
All	7,803	6,008	77%

Of the 7,803 spoken languages of the frame, 1,795 have no entry in the catalog. At the first search, for 595 of them the accessible descriptions were silent on stress, for 456 a description existed but could not be reached, and for 726 no descriptive source was known. The other 18 have a second-hand statement only (5), a statement that describes a neighboring language rather than the language itself (2), no source list (3), or sources not yet read (8). For each language without an entry but with a source list, the bibliography lists the works known and what is lacking.

4.6 Sources for each language

The bibliography lists each language's candidate sources (Section 3.1). For each work it records whether the stress pattern is stated and where, whether the work was opened and found silent, and how a copy can be obtained where that is known. The full release lists 160,673 works for 8,600 Glottolog languoids (Hammarström et al. 2025; sign languages, pidgins and provisional ones included). Of these works, 9,126 are reference pages rather than publications: a language's Glottolog, Wikipedia, Wikidata, Ethnologue or ISO 639-3 page, or an archive or database record. The counts below are restricted to the 7,803 spoken languages of the frame, of which 7,800 have a list, with 151,804 works, reference pages included.

PRELIMINARY

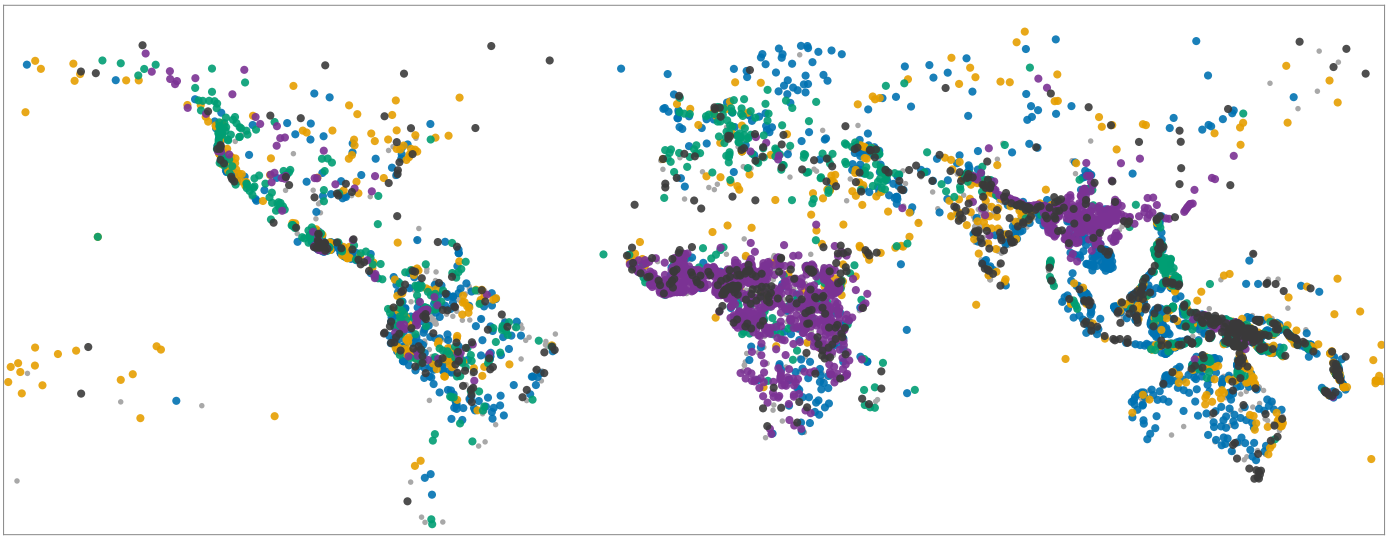

Figure 1: The 7,803 spoken languages of the frame at their Glottolog coordinates, by the stress type of their catalog entry, release of 4 October 2026. Of these languages 6,008 have an entry and 1,795 none; the 82 without coordinates in Glottolog are not shown. Weight-sensitive stress includes the entries stated with hedges, and the fifth class covers no word stress and no type assigned. The legend counts frame languages, not entries.

A language's list starts from the descriptive and phonological works in Glottolog's bibliography and the works cited in catalog entries and in the earlier databases. Works found during the reading itself are added and make up the largest share. Works on tone are gathered in the same way. Duplicate citations are matched, approximately, by first author, year and the opening words of the title, so a work cited in two forms may appear twice. The median list holds 15 works and the longest 1,096.

Each work carries a tag for what it is known to say about stress (for 17,282 works, about tone) and an access class (Section 3.1) with the detail behind it: an open link, a shelf mark or a sign-in page. The list notes whether a copy is on file and, for an opened work, its bearing on the catalog coding: *confirms*, *refines*, *contradicts*, *silent*, or *new* where the language had no entry when the work was read. For a work that states the pattern it gives the passage and page, and says whether the quotation has been checked at that page. Example (5) shows two of the works listed for Kashaya, abridged. A third, classed *scan and deliver*, is Buckley's 1994 monograph, held in print at Harvard's Widener Library.

(5) Kashaya [kash1280] (Pomoan; Buckley 2021: 9–10)

a. *states stress (quoted in catalog).* Buckley, E. 2021 [2023]. M-Words, P-Words, and Accent Phrases in Kashaya. In Bogomolets & van der Hulst (eds.), Word Prominence in Languages with Complex Morphologies, ch. 11. OUP. Digital, open: a copy on file. Sec. 11.2, author's copy pp. 9–10: "Feet are iambic, i.e. right-headed and quantity-sensitive, and are constructed from the left edge of the domain [. . .] The leftmost foot is normally the

PRELIMINARY

Table 5: Works on the frame languages by what is known of their content.

Reading status	Works
quoted in an entry as its source[a]	1,689
domain or secondary-stress passage quoted	556
first-hand statement, language then without an entry	3,938
read: confirms, refines or contradicts an entry[b]	9,737
read: mentions stress, no deciding statement	10,163
read: silent on stress	42,457
read or consulted otherwise[c]	1,995
read for tone: states the tone system (6,834), mentions tone (1,566) or is silent on it (1,969)	10,369
unread: judged likely to state stress (27,322), to describe tone (6,913) or to mention either (1,162), not yet obtained	35,397
unread: seen as an abstract or library record only	6,712
unread: could not be opened	1,626
unread: Glottolog descriptive entry	23,262
unread: Glottolog phonological entry	3,361
unread: cited in our first survey, not yet opened	542
All	151,804

Notes: [a] Sources quoted by the entries that were in the catalog at the re-reading (Section 5.1); the sources of entries made later are counted by their own reading status. [b] The three counts are given separately in Table 7. [c] Studies of morphology read for another purpose, works consulted for an entry, and works we obtained ourselves.

location of primary stress [. . .]"

b. *read: refines the catalog coding.* Oswalt, R. L. 1961. A Kashaya Grammar (Southwestern Pomo). Ph.D. dissertation, Berkeley. Digital, open: a copy on file. Pp. 28, 35, 83–86: "Raised Tone is suprasegmental; it occurs simultaneously with some one syllabic nucleus and its domain is restricted to that one syllable [. . .]" Note: the only reference grammar.

Table 5 classifies every work by reading status. The works quoted in catalog entries as their sources head it. The first-hand statements of its third row are of stress, or of tone alone, for a language that had no entry when the work was read. The two largest unread classes are works judged likely to state the stress pattern or tone system, or to mention either, but not yet obtained, and descriptive works from Glottolog's bibliography that have not been opened. Nearly all works in the first of these classes carry a note saying why they are wanted.

Table 6 gives the works by access class; the library holdings counted are Harvard's. Unreachable open links count as digital, not open, and Glottolog entries not yet looked up as unknown unless Glottolog gives an open link. For the 45,758 unread works with no copy on file, no link or DOI and no record in Harvard's catalog, the access class rests on a search for each in publication indexes, open repositories, digital libraries and library catalogs. That search also found open copies of 720 of them, and 706 of these were read. When each language is counted by the best class any of its works reaches, 7,791 of the 7,800 languages have a work with an open copy or an open link recorded. Of the other nine, the best class reached is digital, not open for two, print only for one and unknown for three, and for three the list is empty. An open work need not mention stress, however, and the work that does describe it may not be open, so a language with an open work can still be without an entry. The bibliography is searchable on the site by language, access class, stress type, confidence, family and area.

PRELIMINARY

Table 6: Works on the frame languages by access class.

Access class	Works
digital, open	83,683
digital, not open	21,397
scan and deliver	8,517
print only	9,283
unknown	28,924
All	151,804

5 Validation

Each coding in the catalog was checked by a second reading of the same source, and compared with an earlier database's coding where one exists and with the language's other descriptions. Table 7 lists the checks made on the entries, with what each covered and what it found. The outcome of the second reading, and of the third reading of the entries it corrected, is given in Appendix B.2. After the initial compilation, every entry then in the catalog was read twice more against the cited sources. The tone statements were checked twice more in the same way, for the tone system, the number of tones, the tone-bearing unit, the relation to stress, downstep, the quotation and the variety named. The corrections of the first check were applied before the second, and the catalog incorporates both sets of corrections.

5.1 Entries re-read against their sources

Once the catalog covered 1,883 languages, every entry was re-read against its sources. The deciding quotation was checked against the source text, and the source compared with the language's other descriptions, those listed by Glottolog (Hammarström et al. 2025) or found by search. The quotation was confirmed for more than nine in ten of these languages (Table 7). Separately, a different classification was proposed for fewer than one language in ten, and for a few no source on stress appears to exist.

5.2 Reconciliation with the earlier databases

The catalog began from the fixed-stress entries of two databases, StressTyp2 (Goedemans et al. 2014) and the appendix of Gordon (2002), each read against the sources it cites, at the page where one is given. From StressTyp2 we took every language whose entry gives the primary stress a single dominant code naming a fixed position without reference to syllable weight: initial, second, third, antepenultimate, penultimate or final. Its secondary-stress code, if any, had likewise to name positions or an alternation without reference to weight. Languages entered with two dominant codes, with variation between positions, or with a weight-sensitive, lexical, unbounded or irregular code were not taken. Of the 699 languages in the archived database, 370 meet the rule. From Gordon's appendix we took the 256 of its 262 rows that give one pattern, in syllables, for all of a language's words. Of the 254 languages these rows name, 194 are among the 370 and are entered once, from StressTyp2. The other 60 come from Gordon's survey. With one entry per Glottolog language, the 370 and the 60 give 394 entries, 345 from StressTyp2 and 49 from Gordon's survey. Rows without a Glottolog code were folded into the entry of the language they describe or set

PRELIMINARY

Table 7: The checks made on the entries, with what each covered and the outcome.

Check	Covered	Outcome
Second reading of the same source	every entry	about one entry in three corrected in some field, not necessarily the stress type
Two further readings against the cited sources, after the initial compilation	every entry then in the catalog	corrections to the entries, each applied after the cited page was read again
Re-reading against the sources and the other descriptions	1,883 languages	quotation confirmed for 1,754; separately, a different classification proposed for 171; for 12 no source on stress appears to exist
Tone statements, first check	2,862 entries with a tone statement at the time	184 with an error in one of the fields checked
Tone statements, second check, after those corrections	2,853 entries at that time	163 with an error, most of them in the three fields filled from the passage's own words (the relation to stress, downstep and the tone-bearing unit)
Database codings read against the pages they cite	394 database entries	fixed stress on our reading for 283; where both databases cover a language, their codings agree for 178 of the 194 languages
Domain of the rule re-read	394 database entries	word 362; word and phrase as one rule 20; phrase 5; domain not named 7
Reading of every open work not yet consulted for its language	7,462 works	with the earlier readings, works on frame languages tagged as confirming an entry's coding 4,257, refining it 4,939, contradicting it 541
Rare patterns re-read against their sources	70 entries	of the 23 singletons, 13 as their sources state them, 1 phrase-level, 3 tendencies, 6 with the secondary stress loosely stated; 13 strings that had marked a secondary stress at some word lengths only and 2 further strings corrected

aside, and rows sharing a code were merged.[3] The coding compared is the dominant primary and secondary code, written out as a stress pattern by word length and read against the sources the database lists. Exceptional and minority patterns recorded beside it are not compared. Where a database coding and the cited page differed at both the first and the second reading, the primary source was consulted again and a note attached to the entry (Appendix B).

On our reading, the cited pages describe fixed stress for 283 entries, and every entry read as fixed falls under the dominant-pattern cut. The other 111 are entered, on the same reading, as weight-sensitive (39), morphological (32), lexical (19) or phrase-level only (2), and the remaining 19 in five other classes. These are languages whose descriptions disagree, languages that appear to have no word stress, and those entered as tone or pitch accent with no word-stress statement, as not determinable or as segmentally conditioned. Table 7 gives the agreement between the two databases where both cover a language. Across the whole catalog, 333 entries (146 of them from the two databases) carry a note setting out where our reading of the pattern differs from a database coding. A difference of this kind is between our reading of one cited description and a coding that may rest on other sources or on an analysis of the same data. It is recorded with the passage for the user to weigh, and is not a finding that the database coding is wrong.

The entry for Sinaugoro shows the form of a reconciliation. Both databases code the language

[3]The 256 rows name 254 languages, Gosiute Shoshone and Cayubaba each appearing under two headings. Thirty-eight of the 60 are in StressTyp2 under a coding outside the rule above. Seven of them, for example, have two dominant codes, initial in disyllables and antepenultimate in longer words.

PRELIMINARY

with stress on odd-numbered syllables from the left, the survey as "Odd numbered from L" and StressTyp2 as "I / i2@1L". The grammar both cite gives the forms in (6).

(6) Sinaugoro [sina1266] (Austronesian; Tauberschmidt 1999: 10–11), stress marked [ˈ] before the syllable as in the source:

a. [ˈbai] 'pig', [ˈnu.ma] 'house', [ˈvo.vo.ka] 'plenty' *σ́*, *σ́σ*, *σ́σσ*

b. [ˈɣa.li.ˈva.ta] 'banana' ("the stress on the first syllable is stronger") *σ́σσ̀σ*

c. [ˈɣa.-ma.-ri.ˈɣo.-ni] 'we are coming down' *σ́σσσ̀σ*

The grammar states the rule in words as well: "On one- to three-syllable words stress and high pitch is on the first syllable", and "On words with four syllables the first and penultimate syllable is stressed" (pp. 10–11). Verbs with a one-syllable subject marker, as in (6c), follow "the normal pattern, stress on the first and penultimate syllable" (p. 11). That rule coincides with the coded alternation at four syllables (6b) and differs from it at three and five (6a, 6c). The entry gives both codings with the quotation and page. Table 9 lists further reconciliations in the same form.

A database row gives a stress position, but the grammar it cites may place that position in the word or in the phrase. Some grammars state word and phrase prominence as one rule. Phrase or utterance conditioning is one of the exception classes, and the sources of the database entries were therefore read again for the domain of the rule alone. For most of them the source states a word-level rule (Table 7). For the rest it states word and phrase prominence as one rule, a phrase-level rule, or a position without naming the domain. In the whole catalog (Table 1) the source places stress in the word for 3,929 of the 6,070 entries and names no domain for 1,642.

5.3 Reading every open work

Every open work then in the bibliography and not yet consulted for its language, 7,462 works, was read in the same way (Section 3.2). Over this reading and the earlier ones, works on frame languages that bear on an entry's coding are tagged as confirming it, refining it (weight, secondary stress, exceptions) or contradicting it. The bibliography quotes each contradicting work's passage and gives its place in the work. In these readings a first-hand statement of stress, or of tone alone, was found for 4,350 frame languages and confirmed by the second reading; each of these languages has an entry in the catalog. Forty-four such statements were at first credited to a language that had no entry, and each was read again against Glottolog's definition of that language. Of these 44 statements, 13 describe the language they were credited to, and an entry has been made from each; the other 31 describe a neighboring language or dialect that Glottolog codes separately and that already has an entry. In the bibliography these 31 statements are marked as describing the other language, and the languages first credited with them remain without an entry. Failed open links were tried a second time. Where the link failed again and no other copy was on file, the work counts as *digital, not open*.

5.4 Rare patterns re-read

Of the 1,315 fixed-stress entries whose rule is complete for words of three to eight syllables, secondary stresses included (Table 3), 23 have a pattern no other entry shares (singletons). Every singleton has been re-read against its source, as has every entry whose pattern lies outside the patterns that rankings of the constraints of Gordon (2002) generate (his factorial typology). In all, 70 entries were re-read (Table 7), among them every entry whose rule string was corrected

PRELIMINARY

as a result. Over half of the singletons are as their sources state them. One, Meskwaki (Algic; Goddard 1991: 160), is coded from a contour that the grammar gives for words inside declarative sentences, while the same words in isolation carry penultimate stress. Three are stated as tendencies (Akawaio, Eibela, Mombum), and the other six turn on secondary stresses that their sources state loosely, as optional, or for some word lengths only. Across the entries re-read, 13 rule strings, among them those of Mombum, Tetum, Mori Bawah, Seke of Vanuatu, Bunak and Bunama, had marked a secondary stress at shorter word lengths and not at longer ones. Their sources state the secondary stress for the shorter lengths only, or with a hedge. These strings have been corrected to mark the primary stress alone, and the secondary stress is described in its own field in each source's words. The corrected strings follow the convention of Section 3.3 for a secondary stress that the source does not state as a general rule. Two further strings were corrected. The string for Sanumá now marks the word-initial secondary stress that its source states as a rule for all words of four or more syllables. The string for Suruahá now places a secondary stress on the first syllable and the primary stress on the last, as its source, the fullest first-hand description of the language, states them.

6 Comparison with the earlier databases of stress and tone

Many of the languages in the catalog are also coded in the earlier databases of stress and tone, and where both code a language the two codings can be compared. The surveys and databases of stress were described in Section 2. They are the survey of Hyman (1977), the survey of Gordon (2002), StressTyp (Goedemans & van der Hulst 2009), the Stress Pattern Database of Heinz (2007) and the compilation of Bailey (1995) behind it, StressTyp2 (Goedemans et al. 2014) and the WALS chapters on fixed stress and weight (Goedemans & van der Hulst 2013a,d). Five of the databases with a tone field were described there as well: WALS chapter 13 (Maddieson 2013), LAPSyD (Maddieson et al. 2013), PHOIBLE (Moran & McCloy 2019), the global dataset of Maddieson & Benedict (2024) and ThoTDB (Maslinsky et al. 2025).

Most of the other databases have a tone field and some have a stress field as well. Four of them are worldwide in scope: the World Phonotactics Database (Donohue et al. 2013), the tone and word-length table of Wichmann (2023), the complexity data of Hartmann & Nichols (2025) and the per-language entries of the Stanford Phonology Archive (Crothers et al. 1979). The World Phonotactics Database is compared in two versions, the snapshot of 2014 that circulated after its site went offline and the re-release of Kalyan & Donohue (2026). The regional databases are EURPhon (Nikolaev 2018), Phonotacticon (Joo & Hsu 2025), SAPhon (Michael et al. 2015), ATLAs (Inman et al. 2025) and UraTyp (Norvik et al. 2026). Two tone tables for the languages of China (Liang et al. 2023; Wang et al. 2024), the tables for Sino-Tibetan (Wu et al. 2023), Bantu (Hartmann et al. 2024), Tai (TaiTone; Dockum & Kirby 2024) and Mixtec (Auderset 2025) languages and the Database of Tonogenetic Events (Sæbø et al. 2025) were also compared. The remaining five are smaller or derived tables: the LAPSyD tone export of Dediu (2021), XTone (Hyman et al. 2004), the Tonal Languages of the World site of the Institute for Linguistic Studies in St Petersburg (Institute for Linguistic Studies, Russian Academy of Sciences 2021) and the tone feature of URIEL+ (Khan et al. 2025). The fifth is the Languages of the World database of the Institute of Linguistics in Moscow (Institute of Linguistics, Russian Academy of Sciences, Moscow 2020). Each draws, directly or through one of the others, on the same descriptive literature. Thirty-one of the databases named above could be obtained and matched to the catalog language by language.

PRELIMINARY

Languages were matched by Glottolog code (Hammarström et al. 2025): a code the database gives was used as given, an ISO 639-3 code was resolved through Glottolog, and a dialect code was raised to its language. A row that names its language without a code was matched only through the catalog entry built from it, the StressTyp2 lect that cites it or, for the Stanford Phonology Archive, the PHOIBLE inventory taken from it. A table written from each database's own documentation maps its values onto the catalog's. Only the fields both hold are compared: whether the language has tone, its tone category and count of tones, its stress type, the position of a fixed stress and weight sensitivity. Where a database has several rows for one language (PHOIBLE's inventories, StressTyp2's lects), the language counts as agreeing when the catalog's value is among those given. The catalog's tone category is set from its count of tones: none for no tone, simple for one or two tones, complex for three or more, and marginal for a tone contrast confined to a small part of the lexicon.

One further table, Maddieson's tally of tone by family (Maddieson 2023), could be compared only by family: it gives a share of tonal languages for each of 34 families and no per-language table. Over the 29 of those families that are also in the catalog, the tonal shares of the tally and of the catalog have a Spearman correlation of 0.8731 and differ by 0.0932 on average. StressTyp's data are compared through StressTyp2, into which they were merged. The survey of Hyman (1977) lists its languages by name, without codes, under each dominant stress placement and within each area or family, and the list was not matched to the catalog. The database behind Maddieson (2007) has not been published with a per-language table. The data behind Hammarström & Mannby (forthcoming), Fedden (2026), Yurayong & Szeto (2023) and Brunelle & Kirby (2015) are unpublished or available only on request. Hyman's own catalog of tone systems (Hyman 2009) and the three Sinitic dialect tone databases of Wang Lining (Wang 2022) have the same status. The World Phonotactics snapshot and Wichmann's table are keyed by ISO code, so that languages sharing a code collapse to one. Gordon's survey and Bailey's compilation name languages without a code and were matched only through the catalog's own entries or StressTyp2 lects that cite them. In this way 202 of Gordon's 262 rows were matched, and none of Bailey's 197, though its data entered the Stress Pattern Database. The Stanford Phonology Archive also names its languages without a code and describes stress and tone in prose comments, from which its values were taken by matching keywords.

6.1 Coverage

The catalog holds most of the languages of each database, counted after matching by Glottolog code (Tables 8 and 10). Table 10 gives the full counts for the twenty-one databases it lists, and its notes say why the counts fall below the published ones. The languages the catalog lacks are concentrated in three collections. ThoTDB lists 252 tonal languages the catalog does not hold, 200 of them in Africa. The World Phonotactics snapshot holds 309 that the catalog does not, 185 of them in Papunesia, and its re-release holds 166 that the catalog lacks. Together the thirty-one databases hold 5,452 languages, of which 4,760 are among the catalog's 6,070 distinct Glottolog codes and 692 are not (Africa 233, Papunesia 207, Eurasia 136, Australia 60, North America 26, South America 14). Of the 692, eleven have no macroarea, and five are outside the catalog's frame of languages. In the other direction, 1,310 of the catalog's languages are in none of the thirty-one (Papunesia 375, Eurasia 346, Africa 311, North America 143, South America 113, Australia 21 and one with no macroarea).

PRELIMINARY

Table 8: Agreement between the catalog and eleven of the databases, over the languages both code. Appendix C gives the numbers of languages behind the rates for tone presence, stress type and fixed position, and the same rates for ten more databases.

Database	In the catalog	Tone presence	Stress type	Fixed position	Weight
WALS chapter 13	510	0.8772	–	–	–
LAPSyD	746	0.8546	0.5877	–	0.8640
Maddieson and Benedict	981	0.8659	0.5917	–	–
ThoTDB (tonal list)	2,127	0.8737	–	–	–
PHOIBLE	1,960	0.6877	–	–	–
World Phonotactics 2014	2,897	0.7740	–	–	–
World Phonotactics 2026	2,518	0.8031	0.5469	–	0.8317
WALS chapters 14 and 15	478	–	0.5962	0.9326	0.8889
StressTyp2	546	–	0.6484	0.9513	0.8986
Stress Pattern Database	337	–	0.6202	0.9422	0.8636
Gordon	202	–	0.6931	0.9635	0.8854

Notes: in the catalog: distinct Glottolog languages of the database that the catalog holds. Each rate is the share of the languages both code for that field whose catalog value is among the database's values; tone presence for ThoTDB is over its list of tonal languages; fixed position is over the languages both code with a fixed position. A dash marks a field the database does not code or a rate not computed for it.

6.2 Tone

On whether a language has tone, the catalog agrees with WALS chapter 13 for 0.8772 of the 391 languages both code (Tables 8 and 10). With LAPSyD, with Maddieson and Benedict and with ThoTDB it agrees for more than four fifths. On the tone category, with the moderately complex value of LAPSyD and of Maddieson and Benedict counted as complex, the rates are 0.8347 against WALS chapter 13, 0.8465 against LAPSyD and 0.8438 against Maddieson and Benedict. Where the catalog and WALS give different categories, the commonest case is a language the catalog enters as complex and WALS as simple. In some of these cases the source lists a contour, a downstep or a rarely used third level among the tones, and the catalog's count includes it where WALS codes a two-way contrast. For Kisi [sout2778] the grammar says that "The tonal system of Kisi has two level and two contour tones". It adds that "An unusual [. . .] feature is an extra-high tone. It is used sparingly in the language, functioning in only a few grammatical contexts" (Childs 1995: 13). WALS codes a simple system and the catalog a complex one. Against LAPSyD and Maddieson and Benedict the commonest case is the reverse, a language the catalog enters as simple and the database as complex. On the exact count of tones the catalog and the databases agree less often. Against every database listed in Table 10 that codes the count, except TaiTone, they agree for between a third and three quarters of the languages, the smallest share against PHOIBLE and the largest against Hartmann and Nichols. Against TaiTone they agree for more than nine in ten (it covers Tai languages only). When counts within one tone of each other are taken as agreeing, they agree for half the languages against PHOIBLE and for most in every other database of Table 10 that codes the count.

PHOIBLE's inventories and ThoTDB's public list record the presence of tone but not its absence. A PHOIBLE inventory is built from one source's phoneme inventory and includes tones only when that source listed them; when an inventory has no tone segments, the source may simply have left tone out, so the inventory does not say that the language has no tone. Of the 897 languages that the catalog classes as tonal, marginal or pitch-accent and that have a PHOIBLE inventory,

PRELIMINARY

the share with a tone segment in at least one inventory is 0.5663. Of the 515 languages whose inventories list tone segments and that the catalog holds with a class stating whether the language has tone, the share classed as tonal, marginal or pitch-accent is 0.9864. For Dizi [dizi1235] the inventory, taken from UPSID, has no tone segments, while the source that UPSID cites says that "Dizi is a register-type tone language with three levels of contrastive tone" (Allan 1976: 379). Only the public list of tonal languages was available to us from ThoTDB, so a language absent from the list is not counted as toneless. Of the 2,305 languages the catalog classes as tonal, the marginal and pitch-accent classes excluded, more than three in four are on the list. Of the 2,035 listed languages that the catalog holds with a class stating whether the language has tone (Table 1), the share classed as tonal, marginal or pitch-accent is 0.9597. The list names the other 82 of the 2,035, which the catalog classes as toneless, and 92 more languages the catalog holds without such a class.

The World Phonotactics snapshot has a tone field only. The re-release of Kalyan & Donohue (2026) gives Glottolog codes, a fuller tone block and a stress block. The re-release agrees with the catalog more closely than the snapshot on the presence of tone and on the count (Tables 8 and 10). Its stress block is compared with the catalog in Table 10, in the same columns as the stress databases.

6.3 Stress

The position of a fixed stress and weight sensitivity agree closely between the catalog and the stress databases, and the stress type less so (Table 10). Where both the catalog and a database code a fixed position, the position agrees at 0.9326 against WALS chapter 14 (178 languages) and at 0.9422 or more against the Stress Pattern Database, StressTyp2, Gordon's survey and UraTyp. Weight sensitivity agrees at 0.8889 against WALS chapter 15 and 0.8986 against StressTyp2 over the languages both code for it. The stress type agrees at 0.5962 against WALS chapters 14 and 15, 0.5877 against LAPSyD, 0.5917 against Maddieson and Benedict, 0.6423 against Hartmann and Nichols and 0.5469 against the World Phonotactics re-release. Against the Stanford Phonology Archive, whose entries describe stress in prose, it agrees at 0.3515. WALS, LAPSyD, Maddieson and Benedict, and Hartmann and Nichols code the stress type in two or three values, the re-release distinguishes fixed, weight-sensitive and lexical stress, and the catalog's stress types (Table 2) are finer. Against WALS and StressTyp2 the largest class of disagreement on the stress type consists of languages the catalog enters as morphologically assigned, lexical or weight-sensitive and the database as fixed. On the position, most disagreements are languages of the catalog's non-fixed types that WALS enters at a position. Of the catalog's 259 languages of non-fixed type that WALS codes, 174 (67%) carry the WALS value "no fixed stress". Where a grammar gives a dominant position without a rule, WALS enters the dominant position and the catalog enters no fixed position. Where a grammar gives a rule with a condition, WALS enters fixed stress and the catalog enters weight-sensitive or morphologically assigned stress. Dobel [dobe1238] is an example of the first. WALS codes penultimate stress and the catalog no fixed position. The grammar says that "It is not predictable whether stress will occur on the penultimate or the final syllable of the root, although it is more common on the penultimate syllable" (Hughes & Hughes 1989: 60–61). Ho-Chunk [hoch1243] is an example of the second. WALS codes fixed stress and the catalog weight-sensitive stress. The grammar says that "Hocąk is indeed basically a fixed stress location language in which the primary stress location is determined from the left edge of the word, but syllable weight also plays a role in the determination of the primary stress location" (Helmbrecht 2010: 119).

PRELIMINARY

The catalog's entries that began from StressTyp2 and from Gordon's survey (Section 5.2) are not independent of those databases, so the rates are given separately for them and for the rest. Of the 512 languages the catalog and StressTyp2 both code for the stress type, the 350 whose entries began from StressTyp2 or Gordon's survey agree with StressTyp2 at 0.7343 and the other 162 at 0.4630. Against WALS, whose stress chapters present StressTyp's data, the entries that began from those databases agree at 0.7249 and the others at 0.4779. The catalog began from the fixed-stress entries of those databases, so the two groups in each pair differ in their stress systems as well as in the origin of their entries. Entries rated very confident agree more often than those rated fairly confident. On the stress type against StressTyp2 the rates are 0.8205 (156 languages) and 0.6214 (243), and on the stress position against WALS chapter 14, over all languages both code, 0.8699 (146) and 0.7296 (233).

6.4 Sources for a sample of the disagreements

A sample of the disagreements was read again by the language model, as in Section 3.2. The sample held 40 drawn at random from each of nine databases (StressTyp2, WALS chapter 13, WALS chapters 14 and 15, LAPSyD, Maddieson and Benedict, PHOIBLE, ThoTDB, and the World Phonotactics snapshot and re-release), 360 in all. Each was read from the catalog's source and, where it is open, from the database's. Two readings were made of each, and the 27 on which they disagreed are counted as contested. Of the 360, 191 (53%) are defensible under either coding, 130 (36%) support the catalog's value, 11 (3%) the database's and one neither. The disagreements were attributed to a coding convention 189 times, an error in the database 67, a different source 58, a different variety 18, a source silent on the point 15 and an error in the catalog 13. The share defensible under either coding is largest against WALS chapter 13 (29 of 40), LAPSyD (28) and Maddieson and Benedict (26). The share supporting the catalog's value is largest against PHOIBLE (28 of 40), where the inventory's source omits tone or the inventory lists contours and phonation types as tonemes.

Of the 13 disagreements attributed to an error in the catalog, the database's value is supported on both readings for 11. For one, neither value is supported, and one is contested. For Pattani [patt1248] the catalog had entered no tone where ThoTDB lists the language as tonal, and the source states that "PaTani has two distinctive tones: high and low" (Saxena 1991: 130). The 13 disagreements concern 11 distinct entries. Nine of the 13 were corrected on that account, and the other four were not changed. Pattani was among the nine. The description its entry follows for word stress states that tone is non-phonemic there, but Saxena gives six minimal pairs and rejects that statement, so the language is entered as tonal and the dissenting statement is recorded in its entry.

A few coding conventions account for 189 of the 360 disagreements. A pitch accent counts as tone in WALS chapter 13 and on ThoTDB's list, and as a class of its own in the catalog. For Somali [soma1255] the catalog enters a pitch accent where WALS enters tone. Hyman writes that "a word can have only one H tone" and proposes that "at an abstract level the prosodic system is accentual and non-tonal" (Hyman 1981: 177). A stress that the grammar states as a rule with a morphological or a weight condition is fixed in WALS and StressTyp2 and morphologically assigned or weight-sensitive in the catalog. An example is Coast Tsimshian [nucl1649], where "In most Tsimshian words, the primary stress is on the last syllable. However, if the last syllable is a suffix or a connective [...], the penultimate (next to last) syllable receives the primary stress" (Dunn 1995: 5). An inventory of tone letters counts contours where a grammar counts levels, and,

PRELIMINARY

as with Kisi above, a downstep or a rarely used extra level makes a system complex in the catalog and simple in WALS. None of these is an error in either coding: each is a choice made when the coding scheme was drawn up, and the catalog's entry quotes the passage so that the other choice can be made from it.

7 Conclusions

The catalog reports what published descriptions say about word stress and tone, and the description quoted in each entry remains the evidence for its language. Each coding is therefore a reading of a written description, and it can be no more complete or accurate than that description.

Each entry refers to the variety its source describes, and names that variety where it differs from the Glottolog language. Variation that no description reports is absent from the catalog. Impressionistic statements in a description, about secondary stress in particular, can reflect the describer's expectations as well as the acoustic signal (de Lacy 2014). A mistake shared by all the descriptions of a language is carried into its entry. Where descriptions conflict, none is discarded: the entry is classed *descriptions disagree* or gives the contradicting statement beside its classification. The catalog also cannot supply what a source omits, whether the domain of the rule or the secondary stresses.

The main limit is that the catalog was compiled by a large language model, not by linguists reading every source. We have spot-checked languages we know and cannot attest that every entry is correct; the checks of Section 5 compare one model reading with another or with an earlier database's coding. Every entry does give the source passage, a translation where needed, and the interpretation the coding rests on, so that a reader can judge it. Errors that both readings can share include a rule projected to unstated word lengths, a tendency read as a rule, word and phrase prominence coded as one, and a source matched to the wrong variety.

The surveys of Hyman (1977) and Gordon (2002), StressTyp2 (Goedemans et al. 2014) and the WALS stress chapters (Goedemans & van der Hulst 2013a,b,c,d) made the typology of word stress a quantitative subject, and this catalog continues their work. It covers more languages, as far as the accessible literature allows, and gives the deciding passage, with page or section, for its classifications. Where the catalog and an earlier database both code a language for the presence of tone or for the position of a fixed stress, they agree for most such languages. The two agree less often on the stress type. In a sample of the disagreements, the sources most often support both codings. The catalog will change as more descriptions become openly available and as its users correct it.

PRELIMINARY

Acknowledgements

The catalog reuses the codings and references of StressTyp2 (Goedemans et al. 2014), compiled by Rob Goedemans, Jeffrey Heinz and Harry van der Hulst, and of the survey of Gordon (2002). It also draws on the language list and bibliography of Glottolog (Hammarström et al. 2025) and the stress chapters of WALS (Goedemans & van der Hulst 2013a). We thank the compilers of all of these. Funding for this project was provided by Anthropic. This is not an official Anthropic project and results and views are not endorsed by Anthropic.

AI contribution

The large language model Claude Fable 5.1 (Anthropic) located and read the sources, drafted the entries, and built the website https://accstack.org. The authors are responsible for the content.

Data availability

The catalog and the bibliography are on https://accstack.org, with pages on the reading procedure and on questions and answers. This paper describes the catalog and bibliography release of 4 October 2026 (version 1.0), which will be archived with a DOI when it is deposited. The counts in this paper are those of that release unless the text dates a count to an earlier release or coding. Earlier releases, including that of 18 September 2026, are kept on the site with their data files.

To cite an entry, give the description it quotes together with the catalog and its release. The catalog itself may be cited as: Schwartz, Matthew D., Kevin M. Ryan, Matthew K. Gordon & Joe Pater. 2026. AccStack, release of 4 October 2026 (version 1.0). Available at https://accstack.org.

The full reading instructions are on the site. Corrections may be proposed through the site, naming the language and, where possible, the page.

The catalog reuses StressTyp2 (Goedemans et al. 2014), which is distributed under a CC BY-NC-ND license, for its codings and references. Its text is not reproduced. Gordon (2002) is cited for its appendix, and Glottolog (Hammarström et al. 2025) for the language list and the bibliographic entries taken from it. Quotations are short attributed extracts, and each audio clip is used under the license of the collection it comes from.

A Reading procedure and coding manual

This appendix gives excerpts from the instructions, the closed lists of readings and types, and the rules for the confidence levels. The full procedure is described on https://accstack.org.

A.1 Excerpts from the instructions

In the reading of every open work, the model was given each language's open works not yet consulted, its Glottolog code, name, family and macroarea (Hammarström et al. 2025), its entry and the statements on stress already found. The instructions, otherwise mostly on how to open and search the works, include:

> Goal: open each listed work, find what it says about word stress (accent, prominence, tone if the work says stress is absent or non-contrastive), quote the deciding sentence(s) verbatim with a page, and say how it relates to what the catalog already records. [...] Never invent a quotation or a page: every quote comes from text you opened; if you could not open a work, say so and leave its quote empty.

The second reading was made with the first report in hand, under these instructions:

> For EVERY work with a quotation: [...] find the quoted passage at the stated locator and say whether it is there verbatim; read the surrounding section yourself and say whether you agree with the reading and with the relation to the catalog, correcting from the fixed lists if not; check the work really describes this languoid. For works [...] marked silent or not opened, spot-check at least two of them [...].

PRELIMINARY

A.2 The closed lists

Each work read receives one of these readings: fixed initial, second, third, antepenultimate, penultimate, final, or other fixed position; weight-sensitive; lexical (unpredictable); morphological (root); tone language with no stress statement; tonal with a stress statement; pitch accent; no word stress according to the source; phrase-level only; unclear or contradictory; none found. The language is then assigned one of the twelve values of the stress-type field, counted in Table 2. A language with too little documentation is entered as *not determinable*; one that proves a duplicate or not a distinct language receives no entry. For fixed stress the classes of exception stated by the source are coded as in Section 4.3; Table 1 lists the other fields.

A.3 Rules for the confidence levels

Of the five confidence levels, very confident, fairly confident and not confident are set by rule, except in the 1,971 entries rated by a judged reading (Section 4.4). The rule counts works, not citations: editions, reprints and variant citations of one work are merged, and works by the same first author count as one, as do a work and a restatement of it. Works on other varieties are listed but not counted. Each work's evidence grade (Section 3.4) and whether it restates another work are recorded separately. *Very confident* requires at least two independent works, read first-hand and not restatements of another work, that agree on the type of system and state the same stress rule. A work that states the same placement in more detail counts as agreeing. The level further requires no confirmed contradiction, no database coding that agrees only in part, and no unread work of a kind ranked above every work read. From highest to lowest the kinds are: grammar, phonology or dissertation on the language; article or chapter on its stress, prosody or phonology; sketch, survey chapter, wordlist or comparative work; database or typological coding. *Fairly confident* is set when a statement of the pattern was confirmed, or read in a snippet that shows it, but the conditions for very confident are not all met. In such an entry either no second independent first-hand work states the same rule; the quoted source itself restates another work, which is then named and listed as wanted if unread; the statement was read only in a snippet, the full copy being listed as wanted; or a higher-ranked work is unread. Any entry with such a work unread is marked *best evidence unread*. *Not confident* by rule means that no statement of the recorded pattern, other than a database coding, was confirmed in a saved copy or shown by a snippet. *Moderately confident* and *disputed* are never set by rule. An entry that the rule would otherwise rate very or fairly confident is referred instead to a judged reading of its sources when it has a confirmed contradiction or a StressTyp2 or WALS coding that agrees only in part, or when its own type is *descriptions disagree*, *not determinable* or *variable*. An entry is also referred to a judged reading when a counted work describes the language's stress as not determinable or variable, or proves not to describe the language. In that reading the entry is rated against the five definitions of Section 3.4, with weight given to the sources' authority and independence and to whether any disagreement bears on the placement and type of stress or only on detail. Each judged rating is checked by a second, independent judgment, and where the two judgments differ the more cautious level stands. Dialect differences alone do not count as contradiction: each entry states the variety of the work it quotes, and a description of another variety is quoted among the further sources with its variety named.

PRELIMINARY

B Examples and outcomes of the checks

This appendix gives an excerpt from the reconciliation of database codings with the pages they cite (Table 9) and the outcome of the second readings and re-readings.

B.1 Reconciliation excerpt

Table 9 shows six rows of the reconciliation in the form the catalog uses; a seventh, Sinaugoro, is example (6). Hatam and Koryak, whose cited pages as we read them state no fixed position for word stress, are entered respectively as phrase-level only and as not determinable. Dehu is entered as a language whose descriptions disagree. An acoustic study (Torres & Fletcher 2022) finds no evidence of the initial primary and third-syllable secondary stress that the grammar describes, and suggests instead that prominence in Dehu is phrasal. Ignaciano, Livonian and Yimas are entered as fixed stress with the pattern the cited page states. Where our reading differs from a database coding, the entry gives both in a database-coding note (Table 1).

Table 9: Six rows of the reconciliation: each database's coding (the survey is that of Gordon 2002), our reading of the page it cites, and the source and page.

Language (family)	Glottocode	Survey coding	StressTyp2 coding	Cited page, as we read it	Source: page
Dehu (Austronesian)	dehu1237	Odd numbered from L minus final	I / i2@1L;i2@1L, no 1R	primary stress initial, one secondary on the third syllable only	Tryon 1968: 1–2
Hatam (Hatam-Mansim)	hata1243	Even numbered from L	S / i2@2L	the accent "cannot be stated in terms of word pattern"; no fixed word-level rule	Reesink 1999: 30–32
Ignaciano (Arawakan)	igna1246	Peninitial	S	initial stress in disyllables, second-syllable stress in longer words	Ott & Ott 1983: 6
Koryak (Chukotko-Kamchatkan)	kory1246	Peninitial	S	"a tendency is noted to place stress on the penultimate syllable of the word; thus, stress is not fixed on a particular syllable of the stem" (our translation)	Zhukova 1972: 40–41
Livonian (Uralic)	livv1244	Odd numbered from L	F/F;I	a trisyllabic final foot in odd-length words, final syllable unstressed (*σσσ̀σσ*, not *σσσ̀σσ̀*)	Tuisk 2015: 177–178
Yimas (Lower Sepik)	yima1243	–	I / i2@1L	primary stress on the first syllable and one secondary on the third, in words of four syllables or more	Foley 1991: 75–77

Notes: a dash means not in the survey. Codes as in (2); a semicolon joins two patterns, `2L` is the second syllable and `F/F` the first heavy syllable, otherwise the first syllable. `F` is unbounded: the heavy syllable can lie at any distance from the edge (Hayes 1995). Livonian, which StressTyp2 codes with two patterns, one of them weight-sensitive, is entered from the survey. The survey gives no page, so the page shown is the one the catalog entry cites for secondary stress, in a study that gives the pattern by word length.

PRELIMINARY

B.2 Second readings and re-readings

The second reading corrected about a third of the first readings in some field (stress type, secondary stress, domain, page or variety). These went to a third reading, which usually agreed with the second and rarely with the first. The re-reading described in Section 5.1 covered the 1,883 languages then in the catalog. The quotation was confirmed in the copy on file for 1,754 of them and in part for 42. For one language it was not found at the page cited. For 52 no copy was then on file, and for 34 the entry had as yet no quotation to check.

C Agreement with each database

Table 10 gives, for twenty-one of the thirty-one databases compared, the number of Glottolog languages each holds after matching, the number of those the catalog holds and the agreement on each field the two share. The number of languages behind a rate is given for tone presence, stress type and fixed position.

PRELIMINARY

Table 10: Languages held by each of twenty-one databases, the number of them the catalog holds, and the agreement on each field over the languages both code. Presence of tone agrees for most languages in every database that codes it, and the category of tone in every database that codes the category except Languages of the World, whose category rate rests on only two languages. The stress type agrees for between one half and three quarters of the languages, the position of a fixed stress for nine in ten or more where both code one, and weight sensitivity for more than eight in ten. The exception on the stress type and on weight is the Stanford Phonology Archive, whose entries describe stress in prose. Of the languages a database records as tonal, the catalog classes nine in ten or more as tonal, marginal or pitch-accent, except against ATLAs.

	Languages		Tone presence		Category	Count	Tonal	Stress type		Fixed position		Weight
Database	total	in catalog	n	rate	rate	rate	given database	n	rate	n	rate	rate
WALS chapter 13	514	510	391	0.8772	0.8347	–	0.9615	–	–	–	–	–
WALS chapters 14 and 15	478	478	–	–	–	–	–	478	0.5962	178	0.9326	0.8889
StressTyp2	556	546	–	–	–	–	–	512	0.6484	267	0.9513	0.8986
Gordon	202	202	–	–	–	–	–	202	0.6931	137	0.9635	0.8854
Stress Pattern Database	342	337	–	–	–	–	–	337	0.6202	173	0.9422	0.8636
LAPSyD	756	746	509	0.8546	0.8465	0.6538	0.9754	713	0.5877	–	–	0.8640
Maddieson and Benedict	994	981	686	0.8659	0.8438	–	0.9744	120	0.5917	–	–	–
PHOIBLE	2,079	1,960	1,329	0.6877	0.6253	0.3689	0.9864	–	–	–	–	–
ThoTDB (tonal list)	2,379	2,127	2,035	0.8737	–	–	0.9597	–	–	–	–	–
World Phonotactics 2014	3,206	2,897	1,695	0.7740	–	0.4730	0.9058	–	–	–	–	–
World Phonotactics 2026	2,684	2,518	1,590	0.8031	0.7400	0.5479	0.9131	629	0.5469	–	–	0.8317
Wichmann	1,485	1,423	1,072	0.8293	0.8338	0.4295	0.9778	–	–	–	–	–
Hartmann and Nichols	321	319	221	0.8235	0.8557	0.7273	0.9247	274	0.6423	–	–	–
Stanford Phonology Archive	197	196	151	0.8609	0.8531	0.5000	–	165	0.3515	–	–	0.7308
EURPhon	393	382	77	0.8701	0.8000	0.7000	0.9870	–	–	–	–	–
Phonotacticon	514	492	344	0.8663	0.8634	0.5847	0.9776	–	–	–	–	–
SAPhon	307	299	163	0.7607	–	–	0.9245	–	–	–	–	–
ATLAs	325	316	199	0.7638	0.8659	0.6739	0.7757	309	0.6505	–	–	–
UraTyp	37	37	–	–	–	–	–	37	0.6486	23	1.0000	0.9143
TaiTone	38	34	34	1.0000	1.0000	0.9259	–	–	–	–	–	–
Languages of the World	83	83	5	0.6000	0.5000	0.5000	–	65	0.7385	–	–	0.9412
Union of thirty-one databases	5,452	4,760										

Notes: languages: distinct Glottolog languages in the database after matching, and those among them the catalog holds. The counts fall below the published ones where dialects are raised to their language or rows without a Glottolog language are dropped. Gordon's rows, which carry no Glottolog code, were matched through the catalog entries built from them or the StressTyp2 lects that cite them. n: languages both code for the field; rate: the share of them whose catalog value is among the database's values; the category, count and weight rates are over the smaller sets of languages both code for those fields. Tone category: none, simple, complex or marginal; count: exact agreement on the number of tones. Tonal given database: of the languages the database records as tonal, the share the catalog classes as tonal, marginal or pitch-accent. For PHOIBLE the presence rate counts an inventory without tone segments as toneless. For ThoTDB the presence rate is the share of its listed languages the catalog classes as tonal in the strict sense, and the column headed Tonal adds the marginal and pitch-accent classes. Fixed position: languages both code with a fixed position and the share whose positions agree; weight: agreement on weight sensitivity. A dash marks a field the database does not code or a rate not computed for it. The union row counts all thirty-one databases compared. Six of the ten not listed are the two tone tables for the languages of China, those for Sino-Tibetan, Bantu and Mixtec languages, and the Database of Tonogenetic Events. The other four are smaller or derived tables: the LAPSyD tone export of Dediu (2021), XTone, the Tonal Languages of the World site in St Petersburg and the tone feature of URIEL+.

PRELIMINARY

References

Allan, Edward J. 1976. Dizi. In M. Lionel Bender (ed.), *The Non-Semitic languages of Ethiopia* (Committee on Ethiopian Studies, Occasional Papers Series 5), 377–392. East Lansing: African Studies Center, Michigan State University.

Auderset, Sandra. 2025. Mixtec Sound Change Database 2.0: Integrating tone change. *Journal of Open Humanities Data* 11. 37. doi:10.5334/johd.322. Data release: MixteCaSo 2.0.0, Zenodo, doi 10.5281/zenodo.14980449 (concept doi 10.5281/zenodo.10143870).

Bailey, Todd Mark. 1995. *Nonmetrical constraints on stress.* Minneapolis: University of Minnesota dissertation. http://web.archive.org/web/20090206103145/http://cf.ac.uk/psych/subsites/ssdb/references/index.html. Not seen; cited after the reference list of the author's Stress System Database, Cardiff University.

Brunelle, Marc & James Kirby. 2015. Re-assessing tonal diversity and geographical convergence in Mainland Southeast Asia. In N. J. Enfield & Bernard Comrie (eds.), *Languages of Mainland Southeast Asia: The state of the art*, 82–110. De Gruyter. doi:10.1515/9781501501685-004. The database of 197 Mainland Southeast Asian varieties is available from the authors on request; see also Kirby & Brunelle, Southeast Asian tone in areal perspective, in Hickey (ed.), The Cambridge Handbook of Areal Linguistics, 2017.

Childs, G. Tucker. 1995. *A grammar of Kisi: A Southern Atlantic language* (Mouton Grammar Library 16). Berlin and New York: Mouton de Gruyter. doi:10.1515/9783110810882.

Crothers, John H., James P. Lorentz, Donald A. Sherman & Marilyn M. Vihman. 1979. *Handbook of phonological data from a sample of the world's languages: A report of the Stanford Phonology Archive.* Stanford: Department of Linguistics, Stanford University. Volume 1 scanned per language by the PHOIBLE project in the archive.org collection phoible; its 197 inventories are redistributed in PHOIBLE 2.0.

Dediu, Dan. 2021. Tone and genes: New cross-linguistic data and methods support the weak negative effect of the "derived" allele of ASPM on tone, but not of Microcephalin. *PLoS ONE* 16(6). e0253546. doi:10.1371/journal.pone.0253546. Data and code: Zenodo, doi 10.5281/zenodo.4762169; the LAPSyD tone export compared is its file LAPSyD-Tone-20200330.csv.

Dockum, Rikker & James Kirby. 2024. TaiTone: Tai lexical tone box dataset (version 1.0.0). Zenodo. https://doi.org/10.5281/zenodo.10689845. Dataset, 926 tone boxes from 112 sources.

Donohue, Mark, Rebecca Hetherington, James McElvenny & Virginia Dawson. 2013. World phonotactics database. Department of Linguistics, The Australian National University. http://phonotactics.anu.edu.au. No longer online; a 2014 snapshot is redistributed with Dediu (2021), doi 10.5281/zenodo.4762169.

Dunn, John Asher. 1995. *Sm'algyax: A reference dictionary and grammar for the Coast Tsimshian language.* Seattle: University of Washington Press. Part II, A Reference Grammar for the Coast Tsimshian Language, revises Dunn's 1979 grammar.

Fedden, Sebastian. 2026. Reassessing Papuan tone: a Canonical Typology of prosodic systems across New Guinea. Talk, 16th International Conference of the Association for Linguistic Typology (ALT16), Lyon, 1 July 2026; abstract in the book of abstracts, pp. 71–72. The 290-language sample is unpublished.

Foley, William A. 1991. *The Yimas language of New Guinea.* Stanford: Stanford University Press.

Goddard, Ives. 1991. Observations regarding Fox (Mesquakie) phonology. In William Cowan (ed.),

PRELIMINARY

Papers of the twenty-second Algonquian conference, 157–181. Ottawa: Carleton University.

Goedemans, Rob, Jeffrey Heinz & Harry van der Hulst. 2014. StressTyp2. http://st2.ullet.net. Version 1 archive of April 2015. Print documentation: Goedemans, van der Hulst & Heinz (eds.), The Study of Word Stress and Accent, Cambridge University Press, 2018.

Goedemans, Rob & Harry van der Hulst. 2009. StressTyp: A database for word accentual patterns in the world's languages. In Martin Everaert, Simon Musgrave & Alexis Dimitriadis (eds.), *The use of databases in cross-linguistic studies*, 235–282. Berlin: Mouton de Gruyter. doi:10.1515/9783110198744.235.

Goedemans, Rob & Harry van der Hulst. 2013a. Fixed stress locations. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*, Zenodo. doi:10.5281/zenodo.13950591. WALS Online v2020.4, chapter 14.

Goedemans, Rob & Harry van der Hulst. 2013b. Rhythm types. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*, Zenodo. doi:10.5281/zenodo.13950591. WALS Online v2020.4, chapter 17.

Goedemans, Rob & Harry van der Hulst. 2013c. Weight factors in weight-sensitive stress systems. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*, Zenodo. doi:10.5281/zenodo.13950591. WALS Online v2020.4, chapter 16.

Goedemans, Rob & Harry van der Hulst. 2013d. Weight-sensitive stress. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*, Zenodo. doi:10.5281/zenodo.13950591. WALS Online v2020.4, chapter 15.

Gordon, Matthew. 2002. A factorial typology of quantity-insensitive stress. *Natural Language & Linguistic Theory* 20. 491–552.

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2025. Glottolog 5.2. Max Planck Institute for Evolutionary Anthropology, Leipzig. https://doi.org/10.5281/zenodo.14006617. https://glottolog.org.

Hammarström, Harald & Emil Mannby. forthcoming. On the world-wide distribution of tonal languages. Forthcoming; cited by ThoTDB as the source of grammar-mined tonal statuses for 4,648 languages (Maslinsky & Vydrin, ThoTDB: A database of the tonal languages of the world, poster, OCP22, 2025, doi 10.5281/zenodo.14813771).

Hartmann, Frederik & Johanna Nichols. 2025. Geospatial effects on phonological complexity in the world's languages. *Linguistic Typology* 29(3). 627–670. doi:10.1515/lingty-2023-0077. Data: Zenodo, doi 10.5281/zenodo.15453927 (Appendix 1) and doi 10.5281/zenodo.20611015 (Appendix A).

Hartmann, Frederik, Seán G. Roberts, Paul Valdes & Rebecca Grollemund. 2024. Investigating environmental effects on phonology using diachronic models. *Evolutionary Human Sciences* 6. e8. doi:10.1017/ehs.2023.33. Data: Zenodo, doi 10.5281/zenodo.10458491 (v1.1).

Hayes, Bruce. 1995. *Metrical stress theory: Principles and case studies*. Chicago: University of Chicago Press.

Heinz, Jeffrey. 2007. *Inductive learning of phonotactic patterns*. Los Angeles: UCLA dissertation.

Heinz, Jeffrey. 2009. On the role of locality in learning stress patterns. *Phonology* 26(2). 303–351. doi:10.1017/s0952675709990145.

Helmbrecht, Johannes. 2010. The accentual system of Hocąk. In Jan Wohlgemuth & Michael Cysouw (eds.), *Rara & rarissima: Documenting the fringes of linguistic diversity* (Empirical Approaches to Language Typology 46), 117–144. Berlin: De Gruyter Mouton. doi:10.1515/

PRELIMINARY

9783110228557.117.

Hughes, Jock & Katy Hughes. 1989. A phonology of Dobel. In Wyn D. Laidig (ed.), *Workpapers in Indonesian languages and cultures, volume 7: Maluku*, 43–76. Ambon: Pattimura University and The Summer Institute of Linguistics.

Hyman, Larry M. 1977. On the nature of linguistic stress. In Larry M. Hyman (ed.), *Studies in stress and accent* (Southern California Occasional Papers in Linguistics 4), 37–82. Los Angeles: Department of Linguistics, University of Southern California.

Hyman, Larry M. 1981. Tonal accent in Somali. *Studies in African Linguistics* 12(2). 169–203.

Hyman, Larry M. 2006. Word-prosodic typology. *Phonology* 23(2). 225–257. doi:10.1017/S0952675706000893.

Hyman, Larry M. 2009. How (not) to do phonological typology: The case of pitch-accent. *Language Sciences* 31(2–3). 213–238. doi:10.1016/j.langsci.2008.12.007.

Hyman, Larry M., David Allison & David Mortensen. 2004. XTone: Cross-linguistic tonal database. Department of Linguistics, University of California, Berkeley. http://xtone.linguistics.berkeley.edu. Offline since 17 January 2020; consulted through Wayback Machine captures of 2013 and 2015.

Inman, David, Natalia Chousou-Polydouri, Marine Vuillermet, Kellen Parker van Dam, Shelece Easterday, Françoise Rose, Alena Witzlack-Makarevich, Kevin Baetscher, Oscar Cocaud-Degrève, Anna Graff, Selma Hardegger, Thomas C. Huber, Tai Hong, Diana Krasovskaya, Raphaël Luffroy, Nora Muheim, André Müller, Alexandra Nogina, David Timothy Perrot, Melvin Steiger & Balthasar Bickel. 2025. The Areal Typology of Languages of the Americas (ATLAs) database. *Scientific Data* 12(1). 933. doi:10.1038/s41597-025-05169-4. Data: CLDF release v2025.1, Zenodo, doi 10.5281/zenodo.15227808.

Institute for Linguistic Studies, Russian Academy of Sciences. 2021. Tonal languages of the world. St Petersburg. https://tonal-langs.iling.spb.ru/. Database and online atlas of the project Tonal languages of the world: database and online atlas (Russian Science Foundation grant 20-18-00250, 2020 to 2022, project leader Elena V. Perekhvalskaya); 12 language records; no version or DOI.

Institute of Linguistics, Russian Academy of Sciences, Moscow. 2020. Languages of the World (Iazyki mira) database, desktop version 4 (beta). https://iling-ran.ru/web/ru/news/201008_langworld. Desktop application, fourth beta version, released 8 October 2020; 110 languages of Eurasia; project supervisor Elena Makarova.

Joo, Ian & Yu-Yin Hsu. 2025. Phonotacticon: a cross-linguistic phonotactic database. *Linguistic Typology* 29(2). 405–431. doi:10.1515/lingty-2023-0094. Data: Phonotacticon v1.0, Zenodo, doi 10.5281/zenodo.20374082 (Joo, Hsu & List, 2026).

Kalyan, Siva & Mark Donohue. 2026. World phonology. OSF project uxshj. https://osf.io/uxshj. Data release of 9 September 2026 (Appendices 1 to 5) of the article The typology of phonological inventories, unpublished as of September 2026; no DOI.

Khan, Aditya, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz & En-Shiun Annie Lee. 2025. URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, 6937–6952. Abu Dhabi: Association for Computational Linguistics. ACL Anthology 2025.coling-main.463; arXiv 2409.18472; data URIELPlus v1.3.1, https://github.com/LeeLanguageLab/URIELPlus.

PRELIMINARY

de Lacy, Paul. 2014. Evaluating evidence for stress systems. In Harry van der Hulst (ed.), *Word stress: Theoretical and typological issues*, 149–193. Cambridge University Press.

Liang, Yuzhu, Lining Wang, Søren Wichmann, Quansheng Xia, Shuai Wang, Jun Ding, Tianheng Wang & Qibin Ran. 2023. Languages in China link climate, voice quality, and tone in a causal chain. *Humanities and Social Sciences Communications* 10(1). 453. doi:10.1057/s41599-023-01969-4. Data: https://github.com/EL-CL/SI_Data.

Maddieson, Ian. 2007. Issues of phonological complexity: Statistical analysis of the relationship between syllable structures, segment inventories, and tone contrasts. In Maria-Josep Solé, Patrice Speeter Beddor & Manjari Ohala (eds.), *Experimental approaches to phonology*, 93–103. Oxford: Oxford University Press. doi:10.1093/oso/9780199296675.003.0007. Preprint: UC Berkeley Phonology Lab Annual Report 1 (2005), 259–268, doi 10.5070/P73cm3w6ck.

Maddieson, Ian. 2013. Tone. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*, Zenodo. doi:10.5281/zenodo.13950591. WALS Online v2020.4, chapter 13.

Maddieson, Ian. 2023. Tone is not predominant: tone is not primordial. In Radek Skarnitzl & Jan Volín (eds.), *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS 2023)*, 1901–1905. Prague: International Phonetic Association. Paper ID 303; no DOI; https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2023/full_papers/303.pdf.

Maddieson, Ian & Karl Benedict. 2024. Global linguistic data (version 1.1.2). Zenodo. https://doi.org/10.5281/zenodo.11205300. Described in Maddieson & Benedict, Frontiers in Psychology 14 (2023), article 1200463, doi 10.3389/fpsyg.2023.1200463.

Maddieson, Ian, Sébastien Flavier, Egidio Marsico, Christophe Coupé & François Pellegrino. 2013. LAPSyD: Lyon-Albuquerque Phonological Systems Database. In *Proceedings of Interspeech 2013*, 3022–3026. Lyon. doi:10.21437/Interspeech.2013-660.

Maslinsky, Kirill, Valentin Vydrin & Dmitry Gerasimov. 2025. ThoT Database. https://thot.huma-num.fr/db/. Described in Maslinsky & Vydrin, ThoTDB: A database of the tonal languages of the world, poster, OCP22, Amsterdam, 2025, doi 10.5281/zenodo.14813771.

Michael, Lev, Tammy Stark, Emily Clem & Will Chang. 2015. South American Phonological Inventory Database v2.1.0. Survey of California and Other Indian Languages Digital Resource. Berkeley: University of California. https://linguistics.berkeley.edu/saphon/. Version 2.1.0 of 25 August 2021; data https://github.com/SAPhon-ling/saphon.

Moran, Steven & Daniel McCloy. 2019. PHOIBLE 2.0. Jena: Max Planck Institute for the Science of Human History. https://phoible.org. Data release: Zenodo, doi 10.5281/zenodo.2626687.

Nikolaev, Dmitry. 2018. The Database of Eurasian Phonological Inventories: a research tool for distributional phonological typology. *Linguistics Vanguard* 4(1). 20170050. doi:10.1515/lingvan-2017-0050. Database: EURPhon, https://eurphon.info.

Norvik, Miina, Robert Forkel, Uldis Balodis, Jeremy Bradley, Michael Dunn, Ilya Egorov, Idaliia Feodotova, Terhi Honkola, Atte Huhtala, Yingqi Jing, Markus Juutinen, Olle Kejonen, Gerson Klumpp, Richard Kowalik, Nikolay Kuznetsov, Miika Lehtinen, Helle Metslang, Merit Niinemägi, Karl Pajusalu, Minerva Piha, Przemysław Podleśny, Mihkel Rünkla, Eva Saar, Sirkka Saarinen, Denys Teptiuk, Pire Teras, Marili Tomingas, Tuuli Tuisk, Janek Vaab & Outi Vesakoski. 2026. Uralic typological database: UraTyp (version 2.0). Zenodo. https://doi.org/10.5281/zenodo.19884642. CLDF dataset, GitHub cldf-datasets/uratyp v2.0; con-

PRELIMINARY

cept doi 10.5281/zenodo.5236365; version 1.0 described in Norvik et al., Journal of Uralic Linguistics 1 (2022), 4–42, doi 10.1075/jul.00002.nor.

Ott, Willis & Rebecca Burke de Ott. 1983. *Diccionario ignaciano y castellano*. Cochabamba: ILV/SIL.

Reesink, Ger P. 1999. *A grammar of Hatam, Bird's Head Peninsula, Irian Jaya* (Pacific Linguistics Series C 146). ANU. doi:10.15144/PL-C146.

Sæbø, Lilja Maria, Eitan Grossman & Steven Moran. 2025. Tonogenesis: a diachronic typology. *Diachronica* 42(3–4). 451–478. doi:10.1075/dia.24022.sae. Data: the CLDF dataset tonodb, https://github.com/cldf-datasets/tonodb (commit 6541c05, 2025).

Saxena, Anju. 1991. Tone in PaTani and Central Tibetan: parallel developments? *Linguistics of the Tibeto-Burman Area* 14(1). 129–136.

Tauberschmidt, Gerhard. 1999. *A grammar of Sinaugoro: An Austronesian language of the Central Province of Papua New Guinea* (Pacific Linguistics Series C 143). ANU. doi:10.15144/PL-C143.

Torres, Catalina & Janet Fletcher. 2022. Phrase-level and edge marking in Drehu. *Glossa: a journal of general linguistics* 7(1). 1–32. doi:10.16995/glossa.5845.

Tryon, Darrell T. 1968. *Dehu grammar* (Pacific Linguistics Series B 7). ANU. doi:10.15144/PL-B7.

Tuisk, Tuuli. 2015. The temporal structure of Livonian tri-, tetra- and pentasyllabic words. *Linguistica Uralica* 51(3). 177–195. doi:10.3176/lu.2015.3.02.

van der Hulst, Harry. 2014. The study of word accent and stress: Past, present, and future. In Harry van der Hulst (ed.), *Word stress: Theoretical and typological issues*, 3–55. Cambridge University Press. doi:10.1017/CBO9781139600408.003.

van der Hulst, Harry, Rob Goedemans & Ellen van Zanten (eds.). 2010. *A survey of word accentual patterns in the languages of the world*. Berlin: De Gruyter Mouton. doi:10.1515/9783110198966.

Wang, Lining. 2022. Hanyu fangyan shengdiao yanbian leixing yanjiu [A study of the types of tone change in Chinese dialects]. Project-completion report of 12 April 2022, National Social Science Fund of China project 12CYY009, National Office for Philosophy and Social Sciences. Unpublished: the three tone databases, each covering the 930 dialect points of the Linguistic Atlas of Chinese Dialects, have not been released.

Wang, Shuai, Yuzhu Liang, Tianheng Wang, Wei Huang, Ke Xu, Aleksandr Mitkov, Shuting Yuan, Jun Ding, Lining Wang, Yongdao Zhou, Quansheng Xia & Qibin Ran. 2024. Unraveling the influence of essential climatic factors on the number of tones through an extensive database of languages in China. *Journal of Language Evolution* 9(1–2). 29–48. doi:10.1093/jole/lzae009.

Wichmann, Søren. 2023. Tone and word length across languages. *Frontiers in Psychology* 14. 1128461. doi:10.3389/fpsyg.2023.1128461.

Wozna, Beata & Theresa Wilson. 2005. *Seimat grammar essentials* (Data Papers on Papua New Guinea Languages 48). Ukarumpa: SIL.

Wu, Baihui, Hanzhi Zhang & Menghan Zhang. 2023. Phylogenetic insight into the origin of tones. *Proceedings of the Royal Society B: Biological Sciences* 290(2002). 20230606. doi:10.1098/rspb.2023.0606. Data and R code: figshare collection, doi 10.6084/m9.figshare.c.6700046 (v1, 2023-06-16).

Yurayong, Chingduang & Pui Yiu Szeto. 2023. Clustering and evolution of phonological systems across languages in Coastal East Asia. *Journal of Language and Culture* 42(2). 78–114. No DOI; https://so03.tci-thaijo.org/index.php/JLC/article/view/274308. The 264-variety feature matrix is not released.

PRELIMINARY

Zhukova, Alevtina Nikodimovna. 1972. *Grammatika koriakskogo iazyka. Fonetika i morfologiia.* Leningrad: Nauka.

PRELIMINARY