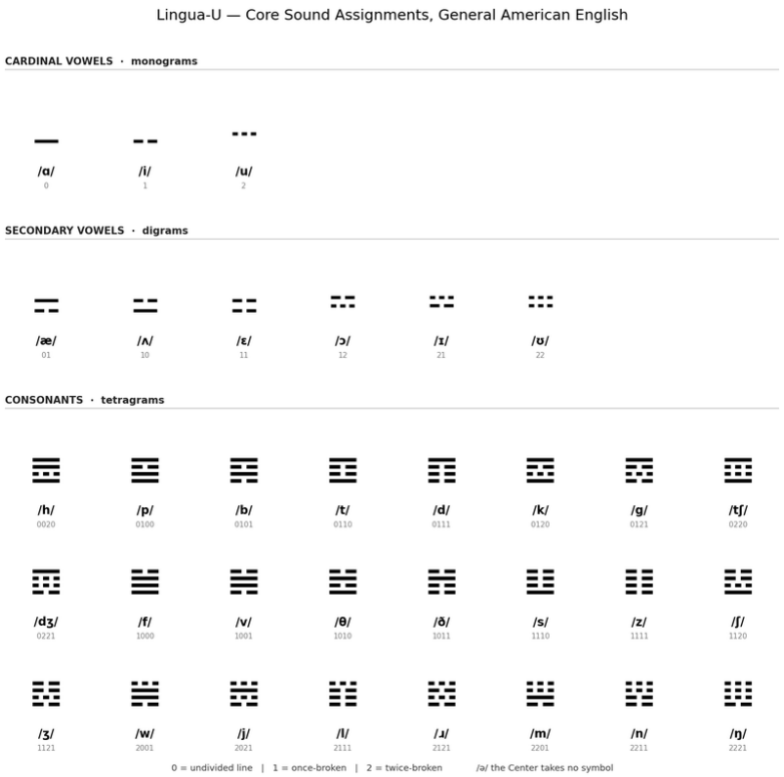
The Lingua-U Whitepaper

A Constructed Ternary Sound-Symbolic System

Version Alpha 1.2



Working draft — not for citation

Status of This Document

This is Version Alpha 1.2. It exists to serve three purposes: to record the current state of the Lingua-U system precisely enough that the full book can be rewritten against it; to stand alone as a coherent statement of what the system is and what it claims; and to be honest about the boundary between what has been established and what remains speculative.

The system described here differs from earlier drafts of Lingua-U chiefly in architecture. The account of what the project is for has changed less than that might suggest.

Version Alpha 1.2 differs from 1.1 in one respect. The cross-linguistic verification that Appendix C listed as outstanding has been carried out. Every cell of the address space has been tested against PHOIBLE, a database of 3,020 phoneme inventories covering 2,177 languages. The results are recorded in Appendix D and folded into Appendix A and the census of Chapter 11. No part of the architecture changed as a result. What changed is that its claims can now be checked.

Three findings were not anticipated. Thirty-nine of the fifty-one open cells have occupants attested in real languages, but only eighteen of those are attested in more than one language in a hundred. Three further cells turn out to be unreachable for the same formal reason as the six already excluded, which the earlier count missed. And two of those six excluded cells have attested occupants after all, which puts the exclusion argument into question rather than settling it.

Two further language attributions surviving in Version 1.1 were found to be wrong, both in Table 9.4. The method used here was designed so that this class of error cannot recur: no language is named from recollection anywhere in this document. Every attribution is generated from a database row and carries its Glottocode.

Those drafts were mixed in register. They stated their aims plainly as poetic, mythic, and spiritual, and most of their claims were appropriately cautious — offered as analogy rather than identity, as images to be felt and lived rather than propositions to be demonstrated. Running alongside that cautious majority was a thinner strand of empirical assertion: that statistical analysis of English vocabulary revealed the correspondences, that systematic evidence supported them.

It is that thinner strand which testing addressed, and did not support. When the archetypal glosses were shuffled within their categories, roughly four-fifths could be justified as fluently as the originals. The supporting word lists were selected rather than exhaustive, and the method admitted no result that would have counted against it. Those particular claims are withdrawn.

Nothing in that result touches the poetic and contemplative aims. A shuffle test can tell you whether a body of evidence discriminates between assignments; it cannot tell you whether a symbolic system is worth inhabiting. Those aims are bracketed here rather than refuted — set aside while the geometry is established, to be resumed in the book with firmer ground beneath them.

What Version Alpha adds is a constructive account of the architecture. Lingua-U is a made thing, generated by stated rules from contemporary phonetic categories, and its correspondences are derived rather than found. This is a weaker claim about the world than the empirical strand attempted, and a stronger claim about the system than that strand could support.

Practice material — chanting, contemplative sequence, devotional application — is omitted here by design. The omission is a matter of sequence, not of standing: the geometry should be stable before anything is built on it.

Appendix C records what is unverified. Readers who want the limitations first should begin there.

Contents

Part I — Position

1. What Lingua-U Is

Traditional sacred alphabets inherit their inventories historically. Hebrew has twenty-two letters because the Phoenician abjad had twenty-two, and the *Sefer Yetzirah*'s cosmology was built to fit that number rather than the number being derived from the cosmology. The Sanskrit varnamala reflects the phonology of an Indo-Aryan language of the first millennium BCE. In each case the inventory came first, as an accident of linguistic and scribal history, and the symbolic architecture was fitted around it afterward.

Lingua-U reverses that order. It begins with contemporary phonetic categories — manner, place, laryngeal state, aperture, orientation — and deliberately generates a symbolic system from them. The inventory is not inherited. It is derived from what the human vocal tract does, as described by the phonetics of the present day, and the symbolic structure is fitted to that description rather than the reverse.

This is the system's distinguishing feature and the source of whatever value it has. It is also the source of its principal limitation, addressed in the following chapter.

At this stage the symbolic system is suggested, not defined. What exists — what has been built, tested, and can be checked — is the alignment of phonemes with a ternary address space. What that alignment means, whether the three values carry stable qualitative character, whether the resulting geometry corresponds to anything a practitioner would recognize: these are open questions. They are treated here as research questions rather than as established results.

The ternary architecture is influenced by the notation of the *Taixuanjing*, Yang Xiong's first-century divinatory text, which employs a base-three system of four-line figures. Lingua-U borrows the notation. It does not borrow the cosmology, the correlative metaphysics, the calendrical scheme, or the divinatory apparatus. Chapter 14 addresses this borrowing in detail.

A comparison, and its limits

Readers meeting a grid of sounds arranged by measurable properties will think of the periodic table. The comparison is worth making, and worth bounding.

Three things carry across. Both systems set aside an inherited list in favor of an arrangement generated from measured properties: Mendeleev abandoned the discovery order of the elements as Lingua-U abandons the alphabet. In both, position predicts properties, so that neighbors in the grid resemble one another in ways the grid was not told about. And in both, the empty cells are the interesting part. Mendeleev left gaps and specified what should fill them; gallium, scandium, and germanium arrived within fifteen years and settled the question. Lingua-U's fifty-one open cells invite the same test, and Appendix C is candid that the test has not been run.

Two things do not carry across. The periodicity Mendeleev found was a real discontinuity in nature, later explained by electron shell structure; his table was a discovery with a mechanism underneath it. Lingua-U's three-way divisions are cuts imposed on continuous articulatory space by decision — eight places of articulation collapsed into three because three is the number the architecture uses. And elements are the same everywhere, whereas the phonemes of General American are a local and

recent arrangement; a Lingua-U built on Hindi or Yoruba would be a different system. Nothing in chemistry corresponds to a dialect.

A third difference is usually described as a further shortfall, and it is not. It is a change of subject. The periodic table says nothing about what sodium signifies, and has no need to; chemistry is not in the business of meaning. Lingua-U is. No arrangement of articulatory features will settle what a sound means, and it does not follow that the question is therefore unsettled or arbitrary. Meaning is made and tested in other workshops — in poems, in liturgies, in songs, in disciplines of attention — and those workshops have standards of their own, older than the ones invoked in Appendix B and not answerable to them. A metaphor is not a hypothesis that failed to reach significance.

So the comparison illuminates one layer of this project and is silent about the other. The architecture of Parts II and III is the kind of thing that can be checked the way a periodic table is checked. The symbolic layer is not waiting on that verdict. It is waiting on use.

That is the invitation this document means to carry, and it is plural. The open cells ask for phonologists and a database. The provisional readings of Chapter 12 ask for something else: for readers willing to sound a phoneme and attend to what it does, to build a practice on the grid and find whether it bears weight, to set a reading against their own experience and report that it is false. A shuffle test is one instrument among several. It answered the question it was built for and is mute on most of the others.

2. What Lingua-U Does Not Claim

Because the earlier drafts overclaimed, this chapter is placed early and stated plainly.

Lingua-U does not claim that the sound-meaning correspondences it describes are universal. The system is built on General American English. Its consonant inventory, its vowel inventory, and the feature analysis underlying both are specific to that variety. Extension to other languages is possible in principle — the address space has room — but has not been carried out.

Lingua-U does not claim that its correspondences were discovered in English vocabulary. An earlier draft argued in places that clusters of meaning around particular sounds revealed archetypal structure, though it more often offered them as poetic images and said so. It was the stronger version that was tested, by shuffling the assignments within their categories and attempting to justify the shuffled versions. Roughly four-fifths of them could be justified as fluently as the originals, and several of the shuffled pairings were better fits than the ones they replaced. The word lists in that draft were selected, not exhaustive, and the method had no failure condition. It has been discarded. Appendix B records the test.

Lingua-U does not claim that the values 0, 1, and 2 carry established qualitative meanings. Provisional readings are offered in Chapter 12, clearly marked as provisional.

Lingua-U does not claim to interpret the *Taixuanjing*. It uses that text's glyph inventory as notation. No reading of Yang Xiong is offered, defended, or implied.

The reference inventory

General American conceals several choices, and Lingua-U fixes them as follows. The reference variety is rhotic. It is treated as unmerged in the cot–caught distinction, so /ɔ/ is contrastive and occupies its own digram; readers whose variety is merged may read that cell as vacant. The vowel /ɒ/

is not part of the inventory — it is listed among the open digrams as a Received Pronunciation value, not an American one. The consonant /ʍ/ is excluded, being moribund in most American speech, as is /x/, which occurs only in loanwords. The glottal stop is excluded as an allophone of /t/ rather than a phoneme. This yields twenty-four consonants and, with the Center unaddressed, nine vowel positions of which six are filled.

This is an idealized inventory rather than a description of any speaker. Lingua-U employs broad phonemic categories as conventional representatives of regions of articulatory space.

What Lingua-U does claim in the register of evidence is narrow and checkable: that the phonemes of General American English can be assigned unique addresses in a ternary space by rules stated in advance; that those addresses preserve phonetic similarity substantially better than chance; that the resulting structure generates open cells which make testable cross-linguistic predictions, not yet verified; and that this constitutes a more defensible foundation for a sound-symbolic practice than an inherited inventory with fitted meanings.

These are the claims that can be adjudicated by test, and they are listed here because a document of this kind should be clear about which of its statements are of that kind. They are not the whole of what the project hopes for. The aims named in the preceding section — poetic, contemplative, and practical — answer to other standards, and the list above should not be read as conceding that nothing else is worth claiming.

3. The Ground: Embodiment and Phonosemantics

Human speech has non-arbitrary dimensions. This is not a fringe position. The association between high front vowels and smallness, documented by Sapir in 1929, replicates across many languages. The correspondence between rounded shapes and sonorant-heavy nonsense words, and between angular shapes and obstruent-heavy ones, is robust enough to have become a standard demonstration. Large-scale corpus work has found sound-meaning associations holding across thousands of languages at rates well above chance.

These effects are real, and they are modest. They are statistical tendencies, not laws. Individual counterexamples are abundant and do not refute them; nor do they license the stronger claim that any particular word's sound explains its meaning.

Traditional sacred-sound systems explored these dimensions long before they were measured, through culturally specific symbolic structures — the mantra traditions of South Asia, the letter mysticism of Kabbalah, the sound-gesture correspondences of eurythmy. Each of these encodes real observations about what speech does in the body. Each also encodes the particular cosmology of its origin, inseparably.

Lingua-U is a contemporary constructive system that approaches the same dimensions through four means: phonetics, which describes what the vocal tract does; empirical phonosemantics, which measures what sound-meaning associations actually hold; phenomenological practice, which attends to what sounds are like from the inside; and ternary symbolism, which supplies a formal structure for organizing the results.

The word constructive is doing real work. Lingua-U does not claim to recover a hidden order. It builds one, from materials that are themselves non-arbitrary, and asks whether the result is useful.

4. Why Three

The ternary architecture is not adopted for numerological reasons, and it does not depend on the *Taixuanjing* for its justification.

The vowels /i/, /ɑ/, and /u/ occupy the three extreme points of the acoustic space available to the human vocal tract. They are maximally dispersed: /i/ high and front, /u/ high and back, /ɑ/ low. Three-vowel systems built on exactly these points recur across unrelated language families. Two-vowel systems are vanishingly rare. Where a language has few vowels, it tends strongly to have these three.

This is a fact about the vocal tract, not about any cosmology. The ternary order in Lingua-U is grounded there. That the *Taixuanjing* also employs a base-three structure is a convergence worth noting and worth nothing more.

The choice of four positions rather than three is likewise forced rather than chosen, though the constraint needs stating precisely. Separating the twenty-four consonants of General American by independently interpretable phonetic features requires four such features. No three-position arrangement built from three single features gave every consonant a unique address; the best of them stranded nine of twenty-four.

This is a constraint on interpretable codes, not on codes in general. An unconstrained optimization can fit twenty-four consonants into twenty-seven cells without collision, and one was found. But its positions are mixtures of features rather than single features, so one cannot say what any position means, and the compositional reading described in Chapter 5 becomes impossible. Four positions are required to give unique addresses while each position remains a single phonetic feature. Appendix B records both searches.

Part II — The Frame

5. Address

Every sound in Lingua-U has an address: a string of ternary digits. Its meaning is read off that address rather than assigned to it.

This is the system's central mechanism and the reason it can make claims the earlier draft could not. When meaning is assigned to a sound directly, any assignment can be justified after the fact, and the justification carries no information. When meaning is derived positionally, exchanging two assignments visibly breaks the derivation — one would be describing a labial sound as though it were articulated at the velum. The system is resistant to arbitrary rearrangement by construction, not by assertion.

Two worked derivations

The consonant /m/ has the address 2201. Read position by position: 2 at the domain gives open resonance; 2 at sub-manner, within that domain, gives nasal closure; 0 at place gives labial, the outermost point of articulation; 1 at the laryngeal position gives modal voice. The composed reading is a continuous voiced resonance, sealed by full closure, sounded at the outer threshold of the body.

The consonant /s/ has the address 1110. Here 1 at the domain gives sustained friction; 1 at sub-manner gives sibilance, friction at its most concentrated; 1 at place gives coronal, the central point of articulate contact; 0 at the laryngeal position gives voicelessness. The composed reading is a concentrated constriction at the center of the mouth, sounded on breath alone.

Exchange those two readings and the derivation fails visibly: one would be describing a labial segment as coronal and a voiced one as voiceless. Nothing obliges a reader to find either reading evocative, but neither can be moved without contradiction.

One qualification carries through the whole system. Because position II is domain-relative, as Chapter 6 sets out, the reading of a digit at position II and beyond is conditional on the digits preceding it: a 2 at sub-manner means nasal within the third domain and affricate within the first. Lingua-U composes along a branching path rather than across four independent axes.

Three tiers of address are used, distinguished by length:

Length	Tier	Holds	Example
1 digit	Monogram	cardinal vowels	0 = /ɑ/
2 digits	Digram	secondary vowels	01 = /æ/
4 digits	Tetragram	consonants	0100 = /p/

Table 5.1 — The three address tiers.

Addresses are always written zero-padded to their tier length, so length alone identifies the tier. The string 01 can only be a digram; 0100 can only be a tetragram. This also distinguishes addresses from ordinary numbers in running prose: 0100 never reads as one hundred.

The digit /ə/ — the schwa — has no address at all. It is the Center, and its lack of an address is discussed in Chapter 8.

6. The Four Positions

The four positions of a consonant address correspond to four dimensions of articulation. They are named rather than numbered in prose — the domain position, the sub-manner position, the place position, the laryngeal position — and referred to as I through IV in tables.

Position I — Domain

Value	Class	Character
0	stops, affricates, and /h/	closure and release; the sound is an event
1	fricatives	sustained constriction; the sound is a channel
2	nasals, liquids, glides	open resonance; the sound is a continuum

Table 6.1 — Position I, the domain.

The domain is the coarsest cut, distinguishing sounds by what fundamentally happens in the vocal tract. Affricates are grouped with the stops rather than the fricatives: they are phonologically non-continuant, stops with a delayed and fricated release, and they pattern with stops in many languages. This grouping also has a structural consequence taken up below.

Position II — Sub-manner

Value	within domain 0	within domain 1	within domain 2
0	unobstructed breath	non-sibilant	glide
1	stop	sibilant	liquid
2	affricate	lateral fricative	nasal

Table 6.2 — Position II, sub-manner. Values are read relative to the domain.

Position II is domain-relative. Its values do not carry the same articulatory meaning across all three domains; a value of 1 means a stop within the first domain and a liquid within the third. This is a genuine feature of the architecture rather than an oversight, and it has an important consequence for how the system should be understood. Lingua-U is better described as a ternary decision tree than as four independent coordinates: each successive position differentiates the space generated by the positions before it. Compositional readings must therefore be conditional on the preceding digits.

The lateral fricative slot at 1-2 is empty in English. It holds the lateral fricatives of Welsh, Navajo, and Zulu; all three attributions have been checked against PHOIBLE and all three hold. The voiceless lateral fricative /ɬ/ is recorded in 146 of the 3,020 inventories, which makes 1210 one of the better-attested open cells in the system.

Position III — Place

Value	Region	Character
0	labial, front	at the outer threshold of the body
1	coronal, mid	at the central point of articulate contact

Value	Region	Character
2	dorsal, postalveolar, glottal	inward, toward the throat

Table 6.3 — Position III, place. Eight English places of articulation are collapsed into three; this is a design decision, not a discovery.

Position IV — Laryngeal

Value	State	Attested in
0	voiceless	English and most languages
1	modal voiced	English and most languages
2	marked: breathy or aspirated	Hindi, Marathi, Thai, Zulu — all four verified

Table 6.4 — Position IV, laryngeal state.

An earlier version of this position distinguished voiceless obstruents, voiced obstruents, and sonorants. That arrangement was defective: sonority is a manner property already encoded at Position I, so the position was a binary contrast in ternary dress, and the combinations it could not reach were then incorrectly described as articulatorily impossible. They were merely excluded by the encoding.

The laryngeal reading corrects this. Its third value is the breathy and aspirated series — the Hindi /bʱ dʱ gʱ mʱ nʱ/ set, and the aspirated stops of Thai. A three-way laryngeal contrast is common cross-linguistically. English uses only the first two values, which is why the third is empty throughout the assignment tables and why the address space has substantial room.

The correction reduced the count of genuinely impossible cells from thirty-one to six and increased the open cells from twenty-six to fifty-one, at a cost of 0.008 in topology preservation. The seven sonorants moved: /m/ from 2202 to 2201, and correspondingly for /n/, /ŋ/, /l/, /ɹ/, /w/, and /j/.

7. Notation

The formal system operates on the digits 0, 1, and 2, which claim nothing. Where the qualitative character of the values is discussed, the provisional names of Chapter 12 are used instead. The change in notation marks the change from established to speculative.

Convention	Use
Zero-padded digits	Addresses. Length identifies tier.
Roman numerals I–IV	Positions within an address.
Subscript 3	Base marker where an address is read as an integer: $0100_3 = 9$.
Middle dot ·	Sequence: one sound followed by another. /aɪ/ = 0·21.
Colon :	Fusion: two sounds' qualities in one segment. /y/ = 1:2.
Superscript mark	Modification of a single sound: rhoticity, nasality, length. /ɜˠ/ = 10^{r}.
U+ form	Unicode codepoints.

Table 7.1 — Notational conventions.

The distinction between sequence and fusion is not decorative. Without it, the pair of symbols standing for /i/ and /u/ could mean either the vowel sequence /iu/ or the front rounded vowel /y/, in which /i/'s tongue position and /u/'s lip posture occur simultaneously. Both operations will be needed as the system extends beyond English.

Fusion and modification are also distinct, and the distinction is easy to lose. Fusion combines the qualities of two sounds that each exist independently: /y/ has the tongue position of /i/ and the lip posture of /u/, and both are separately available as sounds. Modification alters a single sound by a property that is not itself a segment. Rhoticity, nasality, and length are of this kind — there is no separate sound that nasality is, only vowels that carry it.

The separator in sequences is mandatory. Without it, 12·21 and the tetragram 1221 are indistinguishable in transcription.

Part III — The Assignments

8. The Center

The schwa, /ə/, takes no symbol and no address. It is the Center of the system.

This is not an evasion. The schwa is the vowel of minimal articulatory commitment — the tongue at rest, the mouth neither open nor closed, neither front nor back. It is what English vowels reduce to when unstressed, which is to say it is what remains when the distinguishing effort is withdrawn. A system that addresses sounds by their articulatory commitments has nothing to say about the sound defined by having none.

There is a practical benefit and an honest caveat. The benefit: in General American, /ə/ and /ʌ/ are close to identical, distinguished chiefly by stress, and many analyses treat them as one phoneme in two environments. Assigning /ʌ/ the mid-central digram and leaving /ə/ unaddressed resolves a real collision. The caveat: this is a convenient resolution of a genuine near-merger, not a discovery about the sound.

The rhotacized vowel /ɝ/, as in bird or work, is likewise not given its own address, but it is not written as a sequence either. An earlier formulation had it as the Center followed by the rhotic consonant, which misapplies the notation of Chapter 7. Rhotacization is simultaneous rather than sequential: the tongue holds its rhotic configuration throughout the vowel rather than moving into it afterward, and there is no moment at which a speaker of bird has produced a vowel and not yet a rhotic.

It is written instead as the mid-central vowel bearing a rhotic modifier, 10^{r}. This has the additional merit of not requiring the Center to participate in a composition, which it cannot do: an unaddressed point has nothing to compose with.

9. Vowels

Three cardinal vowels take monograms. They are the extreme points of the vowel space described in Chapter 4.

Symbol	Address	Vowel	Position
—	0	/ɑ/	open back
--	1	/i/	close front
---	2	/u/	close back

Table 9.1 — Cardinal vowels.

Secondary vowels take digrams. Position I of a vowel digram gives aperture, Position II gives orientation.

Position	Value 0	Values 1 and 2
I — aperture	open	1 mid, 2 close
II — orientation	central	1 front, 2 back

Table 9.2 — The vowel digram positions.

Symbol	Address	Vowel	Position
⚍	01	/æ/	near-open front
⚌	10	/ʌ/	mid central
⚎	11	/ɛ/	open-mid front
⚏	12	/ɔ/	open-mid back
⚎	21	/ɪ/	near-close front
⚏	22	/ʊ/	near-close back

Table 9.3 — Secondary vowels of General American English.

Three of the nine digrams are unoccupied:

Symbol	Address	Position	Status
⚌	00	open central	[a] — not phonemic in GenAm; 2,600 inventories
⚍	02	open back	/ɒ/ — RP LOT; 224 inventories
⚍	20	close central	/ɨ/ — Russian; 487 inventories

Table 9.4 — Unoccupied vowel digrams.

The three vacant digrams were checked against PHOIBLE for Version 1.2, and the result is worth stating plainly, because it cuts against the reference variety rather than for it. The open central cell at 00, empty in General American, holds [a] — which is recorded in 2,600 of 3,020 inventories, or eighty-six per cent. It is the most widely attested vowel in the database after /i/. The close central cell at 20 holds /ɨ/ in 487 inventories. Neither cell is a gap in the world's vowel systems; both are gaps in English.

Two of the language attributions previously given in Table 9.4 did not survive checking. Russian /ɨ/ is confirmed. Turkish is not: its close unrounded vowel is the back /ɯ/, which belongs at 22 rather than 20. Welsh is not confirmed either — the northern variety is often described with a central /ɨ/, but no PHOIBLE inventory for Welsh records it. Both have been removed.

The relationship between the monograms and the digram grid requires care, because two different numbering logics are in play. A monogram's digit is an identity index — it says which cardinal, not where that cardinal sits. The digram digits are coordinates. The two should not be read as the same kind of quantity.

Locating the cardinals within the digram grid makes the relationship visible. In aperture-and-orientation coordinates /ɑ/ falls at 02, /i/ at 21, and /u/ at 22 — the three extreme corners of the space, which is what makes them cardinal. The digrams at those same coordinates hold the corresponding lax vowels: /ɪ/ at 21 and /ʊ/ at 22. The cell at 02 is vacant, since the reference variety has no lax counterpart to /ɑ/.

So the tiers do not describe two different spaces, and neither do they simply coincide. A monogram and a digram at the same coordinates stand to each other as tense to lax. What the monogram's own digit records is only which of the three cardinals it is.

One qualification is necessary. The IPA descriptions above are approximate. A three-by-three grid cannot separate open-mid from close-mid, so General American /ɔ/ and Spanish /o/ would share a cell. Lingua-U employs broad phonemic symbols as conventional representatives of regions of vowel space; their coordinates should not be read as exact canonical IPA articulations.

Diphthongs are not given addresses. They are movements between vowel positions, and they are written as sequences:

Diphthong	Movement	Sequence
/eɪ/	/ɛ/ → /ɪ/	11·21
/oʊ/	/ɔ/ → /ʊ/	12·22
/aɪ/	/ɑ/ → /ɪ/	0·21
/aʊ/	/ɑ/ → /ʊ/	0·22
/ɔɪ/	/ɔ/ → /ɪ/	12·21

Table 9.5 — Diphthongs as sequences. These are idealized Lingua-U trajectories; actual endpoints vary by speaker and dialect.

Writing diphthongs compositionally has a consequence worth noting. In earlier drafts of Lingua-U, /aɪ/ was treated as one of three fundamental vowels, despite being a trajectory rather than a position. Simple states now take symbols and movements take sequences, which is both more coherent and more defensible: /aɪ/ was /iː/ in Middle English, and a system resting on its present value would rest on a sound change of the last six hundred years.

Vowels beyond English extend by fusion rather than by new cells. The front rounded vowel /y/ combines the tongue position of /i/ with the lip posture of /u/, and is written 1:2.

10. Consonants

The twenty-four consonants of General American occupy twenty-four of the eighty-one tetragram cells.

Symbol	Address	Sound	Index	Positions I–IV
䷎	0020	/h/	6	0 · 0 · 2 · 0
䷀	0100	/p/	9	0 · 1 · 0 · 0
䷁	0101	/b/	10	0 · 1 · 0 · 1
䷂	0110	/t/	12	0 · 1 · 1 · 0
䷃	0111	/d/	13	0 · 1 · 1 · 1
䷏	0120	/k/	15	0 · 1 · 2 · 0

Symbol	Address	Sound	Index	Positions I–IV
	0121	/g/	16	0 · 1 · 2 · 1
	0220	/tʃ/	24	0 · 2 · 2 · 0
	0221	/dʒ/	25	0 · 2 · 2 · 1
	1000	/f/	27	1 · 0 · 0 · 0
	1001	/v/	28	1 · 0 · 0 · 1
	1010	/θ/	30	1 · 0 · 1 · 0
	1011	/ð/	31	1 · 0 · 1 · 1
	1110	/s/	39	1 · 1 · 1 · 0
	1111	/z/	40	1 · 1 · 1 · 1
	1120	/ʃ/	42	1 · 1 · 2 · 0
	1121	/ʒ/	43	1 · 1 · 2 · 1
	2001	/w/	55	2 · 0 · 0 · 1
	2021	/j/	61	2 · 0 · 2 · 1
	2111	/l/	67	2 · 1 · 1 · 1
	2121	/ɹ/	70	2 · 1 · 2 · 1
	2201	/m/	73	2 · 2 · 0 · 1
	2211	/n/	76	2 · 2 · 1 · 1
	2221	/ŋ/	79	2 · 2 · 2 · 1

Table 10.1 — Consonant assignments, ordered by address.

Several regularities fall out of the derivation rather than being imposed on it.

Every voiced and voiceless pair differs at Position IV and nowhere else: /p/ 0100 and /b/ 0101, /t/ 0110 and /d/ 0111, /f/ 1000 and /v/ 1001, and so through all eight pairs. A single articulatory change produces a single symbolic change. This was not true of the earlier trigram arrangement, in which the same voicing contrast produced different symbolic changes for different pairs.

The stops line up by place: /p/ 0100, /t/ 0110, /k/ 0120. So do the nasals: /m/ 2201, /n/ 2211, /ŋ/ 2221. Bilabial, alveolar, velar — a progression from the lips inward, expressed as a single advancing digit.

The consonant /ɹ/ is written with the approximant symbol rather than /r/, which conventionally denotes a trill. General American has an approximant.

11. The Empty Cells

Fifty-seven of the eighty-one tetragram cells hold no English consonant. Version Alpha 1.1 divided them into two kinds, open and excluded, gave the counts as fifty-one and six, and offered candidate occupants for the open cells without checking them. Every cell has now been tested against PHOIBLE by the method of Appendix D. The division survives. The counts do not.

Category	Count	Description
Assigned	24	General American English consonants
Attested candidate	39	Open cells with an occupant recorded in PHOIBLE
Coherent but unattested	9	Well formed under the feature grammar; no attested occupant
Formally unreachable	9	Incompatible under the current definitions; six identified in Version 1.1, three found in verification
Total	81	

Table 11.1 — Tetragram census, verified against PHOIBLE.

The thirty-nine attested candidates are not of one quality, and reporting them as a single number would flatter the system. Attestation is reported here as a proportion of the 3,020 inventories in the database.

Band	Threshold	Cells	Representative
Common	at least 10% of inventories	7	0122 /kʰ/, 601 inventories, 19.9%
Attested	between 1% and 10%	11	1210 /l̪/, 146 inventories, 4.8%
Rare	below 1%	21	2022 /jⁿ̪/, one inventory, 0.03%

Table 11.2 — Attestation bands among the thirty-nine filled cells.

Twenty-one of the fifty-one open cells are therefore occupied only by segments that appear in fewer than one language in a hundred, and several by a segment recorded in a single inventory. A cell filled in that way is filled in a weak sense. The honest summary is that eighteen open cells have occupants a phonologist would call ordinary, twenty-one have occupants that are real but marginal, and twelve have none.

The six excluded cells share a single cause. Position II value 0 within the first domain specifies unobstructed breath, while Position III specifies an oral place of articulation. Under the definitions this system uses, those requirements conflict: /h/ counts as unobstructed precisely because it is glottal, having no oral constriction to locate, so an unobstructed breath assigned a labial or coronal place is asking for a place where the grammar has said there is none. The cells 0000, 0001, 0002, 0010, 0011, and 0012 were therefore held to be unreachable.

That argument is now in difficulty at two of the six. PHOIBLE records /ʍ/, the voiceless labial-velar of Scots and of conservative English which, in thirty-nine inventories, and labialised /hʷ/ in sixty-six. Both are coded as unobstructed breath carrying a labial specification, which places them at 0000. The voiced counterpart /ɦʷ/ places at 0002. Whether this refutes the exclusion depends on a question the system has not answered: whether Position III admits a secondary articulation as a place. If labialisation counts as labiality, two of the six cells are occupied and the exclusion argument is wrong as stated. If it does not, they remain empty, and the system owes an account of where it puts the many hundreds of segments that carry secondary articulation. Either way the claim of Version 1.1 was stronger than the evidence allowed, and it is withdrawn to the status of an open question.

Verification also found three cells that should have been counted as unreachable and were not. Position II value 0 within the third domain specifies a glide, which the system distinguishes from a liquid by the absence of coronal contact — this is what places English /ɹ/ at 2121 rather than beside

/j/. A glide at Position III value 1 therefore asks for a coronal articulation with no coronal contact. The cells 2010, 2011, and 2012 are unreachable for the same kind of reason as the original six, no segment in PHOIBLE occupies them, and the count of unreachable cells rises to nine.

Address 0000 is among the contested cells. It carries the Unicode name TETRAGRAM FOR CENTRE, which invites confusion with the vowel Center of Chapter 8. They are unrelated. The coincidence is an artifact of Unicode's ordering and carries no significance.

Nine cells remain coherent under the feature grammar and empty in the database.

Address	Specification	Why it is likely empty
0021	voiced unobstructed breath, glottal, modal	a modally voiced /h/ is breathy by definition; the sound exists at 0022 as /ɦ/
1012	breathy non-sibilant coronal fricative	breathy /θ̤ ð̤/; the interdentals are rare before any laryngeal marking is added
1102	breathy labial sibilant	a labial sibilant is barely possible; 1100 and 1101 are filled only by labialised /sʷ zʷ/
1201	voiced labial lateral fricative	laterality requires tongue contact; a labial lateral is a contradiction rather than a gap
1202	breathy labial lateral fricative	as above
1220	voiceless dorsal or post-alveolar lateral fricative	1221 holds a retroflex lateral fricative in one inventory; the voiceless counterpart is unrecorded
1222	breathy dorsal or post-alveolar lateral fricative	as above
2100	voiceless labial liquid	liquids are tongue articulations; the labial column of the liquid row is all but empty
2102	breathy labial liquid	as above

Table 11.3 — Coherent but unattested cells.

Chapter 11 of Version 1.1 said that the open cells constitute the system's only predictive content, and that if the architecture is sound they should be fillable by real segments rather than remaining permanently vacant. Taken as a prediction, the result is mixed. Three-quarters of the open cells can be filled; a quarter cannot; and the fills thin sharply toward the edges of the space.

The pattern of failure is the more interesting result, because it is not random. Cells fail in two places. They fail where a place of articulation is incompatible with a manner — labial laterals, labial liquids, labial sibilants — and they fail where a laryngeal setting is incompatible with a manner, which is why the breathy column empties out at the margins. Both are the signature of a decision tree whose positions encode features that are not in fact independent of one another. The system has been describing this as a design decision at Position II. The verification suggests it is a property of all four.

Part IV — The Symbolic Layer

12. Provisional Readings

Everything in this chapter is provisional. Nothing elsewhere in the system depends on it.

The formal system requires no names for its three values. It operates on digits. But a system intended for contemplative use will not be operated on digits alone, and provisional names have been adopted: Hahn for 0, Sheen for 1, Yun for 2.

These names have two justifications, both structural rather than semantic. Each contains its own cardinal vowel: Hahn carries /ɑ/, Sheen carries /i/, Yun carries /u/. And each opens with a consonant drawn from its own domain: /h/ is a domain-0 consonant, /ʃ/ a domain-1 consonant, /j/ a domain-2 consonant. The names demonstrate the system they name. All three close on /n/, a domain-2 consonant, giving them a common grounding.

Their qualitative character is another matter. Working readings have been used — 0 as expansion and release, 1 as refinement and constriction, 2 as grounding and continuity — and these are not baseless: they follow the articulatory character of the domains at Position I. Whether they hold at Positions II, III, and IV, whether they carry any phenomenological weight, and whether practitioners converge on them are open questions.

The temptation at this stage is to elaborate: to assign each of the eighty-one tetragrams a name and a meaning, and to build a correspondence system. That temptation is being resisted deliberately. The earlier draft of Lingua-U did exactly this, and the resulting assignments did not survive testing. The geometry should be stable before anything is built on it.

13. Kindred Systems

Several traditions employ ternary structures that invite comparison. The comparisons are offered as analogy and carry no authority within Lingua-U.

Tradition	First	Second	Third
Sāṃkhya guṇas	rajas — activity	sattva — clarity	tamas — mass, inertia
Peirce	Firstness — quality	Secondness — reaction	Thirdness — mediation
Taixuanjing	天 heaven	地 earth	人— human
Lingua-U	0 — release	1 — constriction	2 — resonance

Table 13.1 — Ternary structures in other systems, offered as analogy.

The guṇas fit more closely than the yin-yang polarity that earlier drafts of Lingua-U invoked. Yin and yang are a binary with a third term appended; the guṇas are natively ternary. One mismatch should be noted rather than hidden: tamas carries negative valence in Sāṃkhya, whereas the grounding principle in Lingua-U does not.

An earlier draft named the three values Yaing, Yin, and Yung, retaining Yin from Chinese cosmology while inventing the other two. That arrangement sat awkwardly with the received sense of the term: in classical yin-yang discourse, yin is conventionally associated with darkness, receptivity, descent, and the earthbound, which corresponds more closely to Lingua-U's third value than to its second. Rather than argue the point, the borrowed term has been dropped.

14. The Glyphs

Lingua-U uses the Tai Xuan Jing symbols encoded in Unicode. The tetragram block runs from U+1D306 to U+1D356 in base-three counting order, with the top line as the most significant digit, so a codepoint follows from an address by arithmetic:

```
codepoint = U+1D306 + address₃
```

This ordering has been verified by rendering the glyphs and measuring their line patterns rather than by assumption. Yang Xiong's own text describes the principle of place value in its eighth chapter, and the line values — undivided, once-broken, twice-broken — correspond to 0, 1, and 2 in the traditional arrangement.

Monograms and digrams require a lookup table rather than a formula. They are split across two Unicode blocks: those containing a twice-broken line were encoded in the Tai Xuan Jing block, while the remainder were unified with existing Yijing symbols, since a stack of undivided lines is visually identical to the ideographs for one, two, and three. The four Yijing digrams follow the Yijing's own bottom-up ordering rather than base-three counting, so their codepoints cannot be derived arithmetically.

Symbol	Address	Codepoint	Unicode name
⚊	0	U+268A	MONOGRAM FOR YANG
⚋	1	U+268B	MONOGRAM FOR YIN
𝌆	2	U+1D300	MONOGRAM FOR EARTH
⚌	00	U+268C	DIGRAM FOR GREATER YANG
⚎	01	U+268E	DIGRAM FOR LESSER YANG
𝌇	02	U+1D301	DIGRAM FOR HEAVENLY EARTH
⚍	10	U+268D	DIGRAM FOR LESSER YIN
⚏	11	U+268F	DIGRAM FOR GREATER YIN
𝌈	12	U+1D302	DIGRAM FOR HUMAN EARTH
𝌉	20	U+1D303	DIGRAM FOR EARTHLY HEAVEN
𝌊	21	U+1D304	DIGRAM FOR EARTHLY HUMAN
𝌋	22	U+1D305	DIGRAM FOR EARTH

Table 14.1 — Monogram and digram lookup. The Unicode names are recorded for reference only.

The Unicode character names have no semantic function within Lingua-U.

This is not merely a disclaimer of convenience. The names are known to be erroneous. Working group document N2988 records that Earth and Human were transposed throughout the block, and the Unicode charts annotate U+1D300 with a note that it is ren, ordinarily associated with human rather than earth. A standard whose own maintainers document its labels as mistaken cannot be treated as conferring meaning.

Two consequences follow that readers will notice, and it is better to state them than to hope they do not. The monogram assigned to the second value carries the name MONOGRAM FOR YIN, which accidentally reproduces the very inversion Chapter 13 describes discarding. And the tetragram names, being drawn from Yang Xiong's divinatory sequence, land arbitrarily: /w/ falls on CLOSED MOUTH and /m/ on CLOSURE, which are apt by pure chance, while /p/ falls on DEFECTIVENESS OR DISTORTION, which is not. None of this signifies. Mathematical notation regularly repurposes shapes; the aleph of set theory carries no Kabbalistic content.

A practical note for implementers: these characters lie in the Supplementary Multilingual Plane and are absent from most common typefaces. Four of the nine digrams are missing even from fonts that cover the Tai Xuan Jing block, since they live in Miscellaneous Symbols. Documents intended for distribution should embed the glyphs as images rather than relying on font fallback.

Appendix A — Complete Assignment Table

This appendix lists every address in the system: the Center, three monograms, nine digrams, and eighty-one tetragrams. Cells not occupied by an English phoneme carry the best-attested segment that the feature grammar places there, together with the number of PHOIBLE inventories recording that segment and up to three languages drawn from those inventories. The selection rules are stated in Appendix D. Example languages are illustrative rather than typical: for a segment occurring in two thousand inventories no three languages are representative, and the entry should be read as evidence that the cell is occupied, not as a claim about where.

Glyph	Address	Sound	Status	Inv.	Example languages
—	—	/ə/	the Center	—	unaddressed by design
⚊	0	/ɑ/	cardinal (English)	—	English
⚋	1	/i/	cardinal (English)	—	English
☷	2	/u/	cardinal (English)	—	English
☰	00	/a/	candidate — common	2,600 (86.1%)	Wangan, Kham Tibetan, Gwandara (Nimbia)
☱	01	/æ/	secondary (English)	—	English
☲	02	/ɑ/	candidate — attested	224 (7.4%)	Dutch, Iron Ossetic, Udmurt
☳	10	/ʌ/	secondary (English)	—	English
☴	11	/ɛ/	secondary (English)	—	English
☵	12	/ɔ/	secondary (English)	—	English
☶	20	/ɨ/	candidate — common	487 (16.1%)	Carib, Tundra Nenets, Amharic
☷	21	/ɪ/	secondary (English)	—	English
☷	22	/ʊ/	secondary (English)	—	English

Table A.1 — The Center, the cardinal monograms, and the vowel digrams. Inv. is the number of PHOIBLE inventories recording the segment, with the percentage of the 3,020 total.

Glyph	Address	Sound	Status	Inv.	Example languages
𝌆	0000	—	unreachable	—	contested: /ʍ, hʷ/ classify here
𝌆	0001	—	unreachable	—	—
𝌆	0002	—	unreachable	—	contested: /ɦʷ/ classify here
𝌆	0010	—	unreachable	—	—
𝌆	0011	—	unreachable	—	—
𝌆	0012	—	unreachable	—	—
𝌆	0020	/h/	English	—	English
𝌆	0021	—	open — unattested	—	—

Glyph	Address	Sound	Status	Inv.	Example languages
	0022	/ɦ/	candidate — attested	105 (3.5%)	Dutch, Kharia, Reβe
	0100	/p/	English	—	English
	0101	/b/	English	—	English
	0102	/pʰ/	candidate — common	584 (19.3%)	Iron Ossetic, English, Kham Tibetan
	0110	/t/	English	—	English
	0111	/d/	English	—	English
	0112	/tʰ/	candidate — common	401 (13.3%)	Northeastern Thai, Burushaski, Lao
	0120	/k/	English	—	English
	0121	/g/	English	—	English
	0122	/kʰ/	candidate — common	601 (19.9%)	Laz, Bengali, Burushaski
	0200	/pf/	candidate — attested	40 (1.3%)	German, Beembe, Ngiemboon
	0201	/ŋv/	candidate — attested	46 (1.5%)	Sango, Gu'de, Mambila
	0202	/pfʰ/	candidate — rare	5 (0.2%)	Beembe, Copi, Khezha
	0210	/ts/	candidate — common	661 (21.9%)	Burushaski, Shigatse Tibetan, Georgian
	0211	/dz/	candidate — common	307 (10.2%)	Chechen, Mingrelian, Reβe
	0212	/tsʰ/	candidate — attested	193 (6.4%)	Chechen, Shigatse Tibetan, Hakka Chinese
	0220	/ʧ/	English	—	English
	0221	/ʤ/	English	—	English
	0222	/ʧʰ/	candidate — attested	223 (7.4%)	Hindi, Burmese, Tsimané
	1000	/f/	English	—	English
	1001	/v/	English	—	English
	1002	/v̩/	candidate — rare	4 (0.1%)	Isoko, Parauk, Ronga
	1010	/θ/	English	—	English
	1011	/ð/	English	—	English
	1012	—	open — unattested	—	—
	1020	/x/	candidate — common	550 (18.2%)	Irish, Tundra Nenets, Bulgarian
	1021	/ɣ/	candidate — common	426 (14.1%)	Meadow Mari, Irish, Igbo
	1022	/ɣ̩/	candidate — rare	1 (0.0%)	Xhosa
	1100	/sʷ/	candidate — attested	35 (1.2%)	Lao, Dan, Avar
	1101	/zʷ/	candidate — rare	16 (0.5%)	Kabardian, Dan, Avar
	1102	—	open — unattested	—	—

Glyph	Address	Sound	Status	Inv.	Example languages
	1110	/s/	English	—	English
	1111	/z/	English	—	English
	1112	/sʰ/	candidate — rare	25 (0.8%)	Shan, Burmese, Korean
	1120	/ʃ/	English	—	English
	1121	/ʒ/	English	—	English
	1122	/ʃʰ/	candidate — rare	3 (0.1%)	Barbareño, Baima
	1200	/ɬʷ/	candidate — rare	2 (0.1%)	Avar, Ronga
	1201	—	open — unattested	—	—
	1202	—	open — unattested	—	—
	1210	/ɬ/	candidate — attested	146 (4.8%)	Northern Khanty, Euchee; Yuchi, Nootka
	1211	/ɮ/	candidate — attested	47 (1.6%)	Daba, Margi, Mofu, South
	1212	/ɮ̰/	candidate — rare	1 (0.0%)	Xhosa
	1220	—	open — unattested	—	—
	1221	/l̪/	candidate — rare	1 (0.0%)	Ao
	1222	—	open — unattested	—	—
	2000	/w̥/	candidate — rare	3 (0.1%)	Breton, Sumo, Washo
	2001	/w/	English	—	English
	2002	/w̰/	candidate — rare	6 (0.2%)	Angami, Eastern Hill Balochi, Ikalanga
	2010	—	unreachable	—	—
	2011	—	unreachable	—	—
	2012	—	unreachable	—	—
	2020	/j̪/	candidate — rare	21 (0.7%)	Aleut, Hopi, Lakkia
	2021	/j/	English	—	English
	2022	/jʱ/	candidate — rare	1 (0.0%)	Dhangar
	2100	—	open — unattested	—	—
	2101	/w˗/	candidate — rare	1 (0.0%)	Mandarin Chinese
	2102	—	open — unattested	—	—
	2110	/l̪/	candidate — attested	73 (2.4%)	Dingri Tibetan, Xumi, gSerpa
	2111	/l/	English	—	English
	2112	/l̪/	candidate — rare	22 (0.7%)	Sumi, Bengali, Nepali
	2120	/l̥/	candidate — rare	3 (0.1%)	Iai, Toda, Wakhi
	2121	/ɹ/	English	—	English
	2122	/ɻ̪/	candidate — rare	18 (0.6%)	Sindhi, Kharia, Nepali

Glyph	Address	Sound	Status	Inv.	Example languages
𝌆	2200	/m̥/	candidate — attested	72 (2.4%)	Ndonga (Kolonkadhi), Shixing, Themchen Tibetan
𝌇	2201	/m/	English	—	English
𝌈	2202	/m̬/	candidate — rare	25 (0.8%)	Sumi, Sindhi, Korku
𝌉	2210	/n̥/	candidate — attested	59 (2.0%)	Themchen Tibetan, Hopi, Mah Meri
𝌊	2211	/n/	English	—	English
𝌋	2212	/n̬/	candidate — rare	13 (0.4%)	Sindhi, Sumi, Angami
𝌌	2220	/ŋ̥/	candidate — rare	28 (0.9%)	Aleut, Iai, Mien
𝌍	2221	/ŋ/	English	—	English
𝌎	2222	/n̤/	candidate — rare	5 (0.2%)	Angami, Khezha, Lotha

Table A.2 — The eighty-one consonant tetragrams.

Appendix B — Methodology Record

This appendix records the tests that produced the current architecture. It is included because the credibility of this document rests on the methods rather than on assertion, and because negative results shaped the design as much as positive ones.

B.0 Method

Each of the twenty-four consonants was encoded as a binary vector over thirteen distinctive features. Phonetic distance between two consonants is the count of features on which they differ. Distance between two addresses is Hamming distance — the count of positions at which the ternary digits differ. Topology preservation is the Spearman rank correlation between those two measures across all 276 consonant pairs. Where an optimization is reported it is hill-climbing over assignments, maximizing that correlation by pairwise swaps and relocations to free cells until no single move improves the score; restarts and random seeds are given with each result.

The feature vectors are given below. They are hand-coded, and several assignments are debatable — whether /ɹ/ counts as coronal, how to treat the double articulation of /w/. They are published so that those calls can be disputed. They are also the weakest input to everything that follows; scaling to an external database, as Appendix C describes, would replace them.

Feature	Consonants carrying it
sonorant	m n ŋ l ɹ w j
continuant	h f v θ ð s z ʃ ʒ l ɹ w j
nasal	m n ŋ
lateral	l
delayed release	tʃ dʒ

Feature	Consonants carrying it
voice	b d g v ð z ʒ dʒ m n ŋ l ɹ w j
spread glottis	h
labial	p b m w f v
coronal	t d n l ɹ s z θ ð ʃ ʒ tʃ dʒ
anterior	t d n l s z θ ð
distributed	θ ð ʃ ʒ tʃ dʒ
dorsal	k g ŋ j w
strident	f v s z ʃ ʒ tʃ dʒ

Table B.1 — Distinctive-feature vectors, given as the set of consonants carrying each feature.

B.1 Version chronology

Several topology scores appear below and they refer to different architectures. The sequence was as follows.

Version	Defining features	Score
Trigram, feature-aligned	Three positions, each a single feature	no unique mapping
Trigram, optimized	Three positions, mixed features, hill-climbed	0.674
Tetragram, first	Affricates in the fricative domain; stop voicing double-encoded	0.630
Tetragram, affricates moved	Affricates in the stop domain; voicing isolated at IV	0.591
Tetragram, laryngeal (current)	Position IV redefined as a three-way laryngeal contrast	0.583
Best feature-aligned (rejected)	Occlusion and stridency replacing domain and sub-manner	0.677

Table B.2 — Architectures tested, in order. Only the fifth row describes the system in this document.

The first three rows are superseded. The comparison in B.3 between an optimized trigram and a tetragram scoring 0.630 is therefore against an earlier tetragram, not the current one. The two corrections that lowered the score — moving the affricates, described in Chapter 6, and redefining position IV, described in B.5 — were both adopted for reasons other than topology.

B.2 The shuffle test

The earlier draft assigned archetypal meanings to each phoneme, supported by lists of English words. To test whether those meanings carried information, the glosses were deranged within their categories and the resulting pairings were justified.

This was an exploratory adversarial exercise, not a controlled experiment, and its provenance should be stated plainly. The derangement was generated programmatically under a fixed seed, but the judgement of whether each shuffled pairing could be justified was made by the same party that had

designed the assignments, with full knowledge of the originals. No blind or independent rater was involved. The criterion — whether a phonetic property constrained the gloss, or only a word list did — was stated in advance, but applied by an interested judge.

One consideration cuts in the result's favor. The finding is that justification is cheap, and a judge motivated to defend the original assignments would tend to understate how easily the shuffled ones could be justified. The direction of any bias runs against the conclusion drawn. The softer figure is the count that resisted, which is straightforwardly a judgement call.

Roughly twenty-one of twenty-seven shuffled pairings could be justified without difficulty, several more persuasively than the originals. The six that resisted were all anchored in articulation rather than vocabulary: /h/ cannot be described as embodied, since it involves no articulator contact; /ŋ/ cannot begin an English word and so supports no vocabulary at all.

A second finding emerged unexpectedly. An audit of whether each gloss echoed its phoneme's letter found seven of nine in the first domain — Primal Path for /p/, True Target for /t/, Core Commitment for /k/ — against one of nine and zero of nine in the others. Those glosses had been generated from orthography, not from sound, in a system that is supposed to be pre-orthographic.

Conclusion: the semantic layer of the earlier draft was largely decorative, and the portion that was not was the embodiment argument. This result motivated the compositional method of Chapter 5.

B.3 Trigram versus tetragram

A three-position architecture was tested against the four-position one. Distinctive-feature vectors were built for all twenty-four consonants, pairwise phonetic distances computed, and topology preservation measured as the rank correlation between phonetic distance and Hamming distance in the code.

An optimized trigram, searched by hill-climbing across forty restarts under seed 7 and constrained to injective assignments, achieved a correlation of 0.674 — slightly better than the tetragram of the day, which scored 0.630. It should be said clearly that this trigram did give every consonant a unique address. Three positions are sufficient to distinguish twenty-four consonants; what they are not sufficient for is doing so while remaining interpretable.

Testing each of its positions against every single feature found purities of 0.67, 0.67, and 1.00. Only the third corresponded to a single phonetic feature; the other two were mixtures the optimizer had found useful. Compositional derivation is impossible under such a code, because one cannot say what a position means, and the mechanism of Chapter 5 depends on being able to say so.

The trigram also leaves only three cells free of twenty-seven — eighty-nine per cent occupancy, which forecloses the cross-linguistic extension the architecture is meant to permit. It was rejected on those two grounds rather than on its score.

The four-position architecture was adopted for these reasons rather than for its topology score.

B.4 Feature-aligned optimization

An attempt was made to optimize the tetragram by permuting which feature occupies which position and which value maps to which digit. This was found to be vacuous: Hamming distance is invariant

under both permutations, so neither can change topology preservation. Those choices affect readability and pedagogy, not geometry.

What can change the score is which features are chosen. Ten candidate single-feature ternary partitions were built and all two hundred and ten four-way combinations tested. Only five produce unique addresses for all twenty-four consonants. The highest-scoring, at 0.677, groups nasals with stops and reduces the third domain to three members, which destroys the domain reading; it was rejected. The current architecture scores 0.583.

A random baseline over four hundred feature-blind assignments, seed 3, produced a mean correlation of −0.004 and a maximum of 0.174. Every feature-aligned design is far above chance, which is the strongest quantitative statement this document can make.

B.5 The laryngeal correction

An external review observed that the category described as structurally impossible was overstated. The observation was correct, and pursuing it revealed a deeper defect in Position IV, described in Chapter 6. The correction reduced impossible cells from thirty-one to six, raised open cells from twenty-six to fifty-one, and cost 0.008 in topology preservation. The arrangement remains injective.

Appendix C — Limitations and Open Questions

What follows is not comprehensive, but it records what is known to be unfinished.

- Cross-linguistic candidates for the open cells have now been verified against PHOIBLE; the method and its results are in Appendix D. What remains unfinished is the interpretation: twenty-one of the fifty-one open cells are filled only by segments occurring in fewer than one per cent of inventories, and the system has not decided what threshold of attestation makes a cell genuinely occupied.
- The feature analysis of Appendix B still rests on twenty-four hand-coded vectors. The verification of Appendix D runs on PHOIBLE's coding of 1,072 consonant segments, but the topology measurements — the correlation between phonetic and Hamming distance — have not been recomputed at that scale. Doing so is the strongest remaining test of whether the address space predicts genuine phonetic similarity.
- Position II is domain-relative, so the system is a decision tree rather than a coordinate space. Compositional readings must be conditional on preceding digits. This is stated but not yet worked through for all cells.
- Position III collapses eight English places of articulation into three. The particular grouping is defensible but is a design decision, not a discovery.
- The vowel digram grid cannot separate open-mid from close-mid, so General American /ɔ/ and Spanish /o/ would share a cell.
- The vowel tier has three free digrams against fifty-one free tetragrams. Vowel extension therefore relies on fusion rather than on unoccupied cells, and this mechanism has not been tested at scale.
- The provisional readings of Chapter 12 have no empirical support. Whether practitioners converge on them is untested.
- No pseudoword or rating study has been conducted. This is the obvious next empirical step: whether naive listeners rate nonce words built on domain-2 consonants as more grounding than those built on domain-1 consonants is a cheap and genuinely informative experiment.
- The system is built on General American English and has not been extended to any other variety, let alone any other language.

A final caution, directed at the author as much as the reader. The strongest temptation in a project of this kind is to treat structural elegance as evidence. It is not. The nasals lining up by place is a pleasing fact about the feature analysis, and it would be equally pleasing under an arrangement that meant nothing. The tests in Appendix B exist because the pleasure of a well-formed table is not informative, and the same discipline will be needed for every subsequent version.

Appendix D — Cross-Linguistic Verification

This appendix records how the candidate occupants of Appendix A were produced. It exists because the previous attempt at the same table was produced from recollection and contained errors, and because a reader has no way to tell the two kinds of table apart by looking at them.

D.1 Method

The source is PHOIBLE, which records 3,020 phoneme inventories for 2,177 languages, drawn from several independent survey traditions. Two files were used: the inventory table, which states which segments occur in which doculect, and the segment feature table, which gives each of 2,162 segments a vector of thirty-seven distinctive features.

The verification does not ask which sounds belong in which cell. It implements the four positions of Chapter 6 as a function from a feature vector to an address, and then runs every segment in the database through it. No segment is placed by judgement, and no language is named except by reading a database row. The substantive decisions are in the function, and they are these:

Position	Rule as implemented
I — domain	nasal, or approximant, gives the third domain; continuant gives the second; otherwise the first
II — sub-manner	within the third domain, a lateral or coronal resonant is a liquid and any other is a glide; within the second, the strident feature separates sibilant from non-sibilant; within the first, delayed release separates affricate from stop
III — place	labial takes precedence; then coronal, split by the anterior feature into value 1 and value 2; everything else, including dorsal, pharyngeal, and glottal, takes value 2
IV — laryngeal	spread glottis with periodic voicing gives breathy; spread glottis on an obstruent gives aspirated; spread glottis on a sonorant gives voiceless, since a voiceless nasal is not an aspirated one; otherwise voicing decides

Table D.1 — The classifier, stated so that it can be disagreed with.

One exception is coded explicitly. In the unobstructed-breath branch, glottal spread is constitutive of the segment rather than a laryngeal modification of a closure, so it is not read at Position IV. Without this, /h/ would classify as aspirated and take the address 0022, which is /ɦ/.

D.2 The unit test

A classifier that agrees with the whitepaper because it was tuned to agree proves nothing. The test applied was that the function must reproduce all twenty-four English consonant addresses of Table 10.1 from PHOIBLE's feature coding alone. It does: twenty-four of twenty-four, with no collisions. The vowel digram function reproduces eight of the nine coordinates in Chapter 9.

The two disagreements encountered along the way are worth recording, since both are places where Lingua-U and PHOIBLE describe the same sound differently.

PHOIBLE codes English /ɹ/ as non-consonantal, which is to say as a glide. A liquid/glide split defined by consonantality therefore sends /ɹ/ to 2021, where /j/ already is, and the system loses its injectivity. Defining the liquid class by laterality or coronal contact instead reproduces Chapter 10 exactly. This is a real choice and not a neutral one: it is the classifier agreeing with Lingua-U against one of PHOIBLE's feature assignments.

PHOIBLE codes the symbol ⟨ʌ⟩ as back, which would place /ʌ/ at 12 beside /ɔ/ rather than at 10. Lingua-U's assignment reflects the General American realisation as mid-central, which is the standard description of the modern variety and not a convenience. Something in the neighbourhood corroborates it: the most widely attested occupant of cell 10 across the database is /ə/, in 663 inventories — the sound that Chapter 8 argues is near-merged with /ʌ/ in this variety and left unaddressed for that reason.

D.3 What was found

Of 2,162 feature-coded segments, 1,963 occur in at least one inventory as a non-marginal phoneme. Of those, 631 are vowels or tone-bearing units, 260 carry an airstream or laryngeal property the four positions cannot represent, and 1,072 classify into the eighty-one cells. The census that results is Table 11.1; the cell-by-cell result is Table A.2.

The distribution is severely uneven. The mean occupied cell holds eighteen segments, and the crowding falls where English already is.

Cell	English phoneme	Segments sharing the cell	Examples of what else lands there
0120	/k/	100	ʔ, ʈ, c, q
0121	/g/	77	ɟ, ŋg, ɖ, gʲ
0100	/p/	59	kʷ, kp, pʼ, kʷʼ
2111	/l/	55	r, ɾ, ɭ, ɽ
0101	/b/	53	gb, mb, ɓ, gʷ

Table D.2 — The five most crowded cells.

This is the cost of collapsing eight places into three and every airstream into none. The cell holding /k/ also holds the glottal stop, the retroflex, the palatal, and the uvular stop; the cell holding /p/ also holds the labial-velars, since Position III reads labiality before dorsality. A reader who takes an address to identify a sound is reading it wrongly. An address identifies a region, and English happens to sit in the densest regions of it.

D.4 What the method cannot see

Five limitations bear on how much weight Appendix A can carry.

PHOIBLE records contrastive phonemes in doculects, not the articulable. A cell reported as unattested means that no source in the database records a segment of that description as a phoneme of any language it surveyed. It does not mean the sound cannot be made. It also does not mean much about frequency of use: 3,020 inventories describe 2,177 languages, so some languages are counted several times and well-studied families are over-represented.

Roughly one row in twenty — 4.7 per cent — involves a segment with no entry in the feature table. Those rows were skipped rather than guessed at. PHOIBLE also records about a hundred segments with disjunctive coding, where a source could not decide between two analyses; these were likewise skipped.

The 260 excluded segments are not a rounding error. Clicks, ejectives, and implosives have no position in a four-place address that encodes manner, place, and laryngeal state but not airstream. They do not fail to fit; they fit too well, landing silently on cells already occupied by their pulmonic

counterparts. An ejective /pʼ/ classifies to 0100, which is /p/. Excluding them from candidate selection prevents a false claim of occupancy, but the underlying gap is architectural and is not closed by this verification.

Secondary articulation raises the same question in a milder form. Three cells — 1100, 1101, and 1200 — are filled only by labialised segments, where the labiality that put them there is secondary to an articulation elsewhere in the mouth. The same issue decides the status of 0000 and 0002, as Chapter 11 sets out. Where a cell has a candidate without secondary articulation, that candidate was preferred; where it has none, the cell is filled by a segment whose claim to be there is weaker than its inventory count suggests.

One consequence of that is visible in Appendix A and is worth naming before a reader trips on it. English appears as an example language for cell 0102, which holds /pʰ/, although Chapter 6 states that English uses only the first two laryngeal values. Both are right. Several PHOIBLE sources analyse English with a phonemic aspirated series rather than with allophonic aspiration, and the database records the analysis its sources gave. The reference inventory of Chapter 2 is a choice among available analyses of English, not a fact about English, and the verification makes that visible in a way the earlier tables did not.

Phonemes marked marginal in the source — loan phonology, restricted distribution — were excluded from every count. This is conservative and it moves some cells from rare to unattested.

D.5 Selection rules

Where several segments occupy a cell, the one shown in Appendix A is the segment appearing in the most inventories, with segments carrying an unrepresentable airstream excluded and segments carrying a secondary articulation deprioritised. Example languages are the three languages contributing the most independent source inventories for that segment, with ties broken by how many inventories PHOIBLE holds for the language overall, and with no language cited more than three times across a table so that the examples span families rather than repeating the best-documented one.

Attestation bands are set at ten per cent and one per cent of inventories. These thresholds are conventional and nothing in the architecture requires them; the inventory counts are given in Appendix A so that a reader who prefers a different line can draw it.

D.6 Reproduction

The verification is three short scripts: one that maps a feature vector to an address and unit-tests itself against Table 10.1, one that runs the database through it and emits the census, and one that renders the glyphs and writes this document. They depend on nothing but the two PHOIBLE files, which are public. Any claim in Appendix A can be traced to a row.

Appendix E — References

Version Alpha states a number of claims about phonetics, phonosemantics, and the history of the notation it borrows. Sources are given here so those claims can be checked. This is a working list rather than a complete apparatus, and the sacred-word literature in particular is under-represented — a gap the book will need to close.

Phonosemantics and sound symbolism

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences, 113*(39), 10818–10823.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences, 19*(10), 603–615.

Köhler, W. (1929). *Gestalt psychology*. Liveright.

Ohala, J. J. (1994). The frequency code underlies the sound-symbolic use of voice pitch. In L. Hinton, J. Nichols, & J. J. Ohala (Eds.), *Sound symbolism* (pp. 325–347). Cambridge University Press.

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia: A window into perception, thought and language. *Journal of Consciousness Studies, 8*(12), 3–34.

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology, 12*(3), 225–239.

Phonetics and phonological typology

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. Harper & Row.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Blackwell.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language, 48*(4), 839–862.

Moran, S., & McCloy, D. (Eds.). (2019). *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History. https://phoible.org

The notation

International Organization for Standardization. (2005). *Proposal to correct the names of the Tai Xuan Jing symbols* (ISO/IEC JTC1/SC2/WG2 Document N2988).

Nylan, M. (Trans.). (1993). *The canon of supreme mystery by Yang Hsiung: A translation with commentary of the T'ai Hsüan Ching*. State University of New York Press. (Original work c. 2 BCE)

The Unicode Consortium. (2023). *Tai Xuan Jing symbols: Range 1D300–1D35F*. https://www.unicode.org/charts/PDF/U1D300.pdf

Walters, D. (1983). *The alternative I Ching*. Aquarian Press.

Sacred-word traditions

Beck, G. L. (1993). *Sonic theology: Hinduism and sacred sound*. University of South Carolina Press.

Kaplan, A. (Trans.). (1997). *Sefer Yetzirah: The book of creation* (Rev. ed.). Weiser.

Padoux, A. (1990). *Vāc: The concept of the word in selected Hindu Tantras* (J. Gontier, Trans.). State University of New York Press.

Steiner, R. (1984). *Eurythmy as visible speech* (V. & J. Compton-Burnett, Trans.). Rudolf Steiner Press. (Original lectures delivered 1924)

This section is the thinnest in the document. Chapter 1 makes claims about how the Hebrew and Sanskrit inventories came to have the shapes they have, and those claims deserve better support than a single reference each.